Computational Aspects of Communication Amid Uncertainty
by
Badih Ghazi

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

This thesis focuses on the role of uncertainty in communication and effective (computational) methods to overcome uncertainty. A classical form of uncertainty arises from errors introduced by the communication channel but uncertainty can arise in many other ways if the communicating players do not completely know (or understand) each other. For example, it can occur as mismatches in the shared randomness used by the distributed agents, or as ambiguity in the shared context or goal of the communication. We study many modern models of uncertainty, some of which have been considered in the literature but are not well-understood, while others are introduced in this thesis:

Uncertainty in Shared Randomness

- We study **common randomness and secret key generation**. In common randomness generation, two players are given access to correlated randomness and are required to agree on pure random bits while minimizing communication and maximizing agreement probability. Secret key generation refers to the setup where, in addition, the generated random key is required to be secure against any eavesdropper. These setups are of significant importance in information theory and cryptography. We obtain the first explicit and sample-efficient schemes with the optimal trade-offs between communication, agreement probability and entropy of generated common random bits, in the one-way communication setting.

- We obtain the **first decidability result** for the computational problem of the *non-interactive simulation of joint distributions*, which asks whether two parties can convert independent identically distributed samples from a given source of correlation into another desired form of correlation. This class of problems has been well-studied in information theory and its computational complexity has been wide open.

Uncertainty in Goal of Communication

- We introduce a model for **communication with functional uncertainty**. In this setup, we consider the classical model of communication complexity of Yao, and study how
this complexity changes if the function being computed is not completely known to both players. This forms a mathematical analogue of a natural situation in human communication: Communicating players do not a priori know what the goal of communication is. We design efficient protocols for dealing with uncertainty in this model in a broad setting. Our solution relies on public random coins being shared by the communicating players. We also study the question of relaxing this requirement and present several results answering different aspects of this question.

Uncertainty in Prior Distribution

- We study data compression in a distributed setting where several players observe messages from an unknown distribution, which they wish to encode, communicate and decode. In this setup, we design and analyze a simple, decentralized and efficient protocol.

In this thesis, we study these various forms of uncertainty, and provide novel solutions using tools from various areas of theoretical computer science, information theory and mathematics.

Thesis Supervisor: Ronitt Rubinfeld
Title: Professor of Computer Science, MIT

Thesis Supervisor: Madhu Sudan
Title: Professor of Computer Science, Harvard
Acknowledgments

I have been very fortunate to have great mentors, collaborators, friends and family without whom this work would not have been possible.

I am extremely thankful to Madhu who has been a fantastic advisor throughout my PhD. Madhu’s warm intelligence, encyclopedic knowledge, humility and creativity make it impossible to work with him without admiring him. Not once did I leave a meeting with Madhu without having a new idea or a fresh take on a problem. Among the things that I will miss the most about graduate school are my meetings with him.

I am very grateful to Ronitt who was incredibly helpful on every occasion long before she became my official advisor. I first got the chance to work with Ronitt in my second year at MIT when I TA’ed her wonderful course on Randomness and Computation. Ever since, her support has been phenomenal. I am very fortunate to have been advised by her.

Many thanks to Vinod Vaikuntanathan and T.S. Jayram for agreeing to be on my thesis committee. Vinod’s sharp-sightedness and ability to instantly detect connections between almost any theory question and cryptography are spectacular. I have learned a lot from Jayram’s expertise in communication and information complexity, and I am amazed by his eclectic knowledge that spans so many areas within the theory and applications of computer science.

This thesis has benefitted from discussions and exchanges with several researchers including: Boaz Barak, Mohammad Bavarian, Clément Canonne, Mohsen Ghaffari, Oded Goldreich, Mika Göös, Venkat Guruswami, Brendan Juba, Sudeep Kamath, Swastik Kopparty, Jingbo Liu, Elchanan Mossel and Ilya Razenshteyn.

I would also like to thank Vitaly Feldman for his supervision, along with Jayram, during my internship at IBM Almaden in the summer of 2016 and for his help and advice since then.

Also, many thanks to Venkat Guruswami who hosted me for a great week at CMU during which I started thinking about Common Randomness Generation. I hope to get to collaborate with Venkat in the future!

I would also like to thank Dina Katabi for hosting me during two summers at MIT. Dina’s work ethic and dedication were truly inspiring. I would also like to thank Piotr Indyk for the collaboration during that period and his help since then.

I was fortunate to have other great collaborators including: Haitham Hassanieh, Eric Price, Lixin Shi, Eric Blais, Joshua Brody, Elena Grigorescu, Venkata Gandikota, Pratish Kamath, Elad Haramaty, Prasad Raghavendra, Mohammad Bavarian, Ronald L. Rivest, Ilan Komargodski, Pravesh Kothari, Euiwoong Lee, and Rudiger Urbanke.

In particular, I would like to thank Pratish with whom I have had frequent and fruitful collaborations. Besides being a fantastic person, Pratish’s diligence, mathematical rigor and clarity of thought are truly remarkable.

I had the chance to get to know many friends in the Theory of Computation group at MIT with whom I had numerous interesting conversations: Mohsen Ghaffari, Pratish Kamath, Elad Haramaty, Alan Guo, Ilya Razenshteyn, Jerry Li, Gautam G Kamath, Arturs Backurs, Adam Sealfon, Cameron Musco, Katerina Sotiraki, Maryam Aliakbarpour, Nishanth Dikkala,
Akshay Degwekar, Prashant Vasudevan, Elena Mădălina Persu.. Apologies to those whom I am forgetting!

Also, thanks to the other researchers that I have had the chance to interact with during graduate school (including Guy Kindler, David Woodruff and Ilias Diakonikolas).

I would like to thank the administrative assistants in the Theory of Computation Group at MIT, especially Joanne Hanley, whose help and kindness has shielded us from all the bureaucratic requirements.

Back as an undergraduate student at the American University of Beirut, I was extremely lucky to work with Louay Bazzi. It was Louay’s outstanding courses and his mentorship during my first research project that sparked my interest in theoretical computer science. Also, many thanks to AUB professors Ibrahim Abou-Faycal, Fadi Karameh and Kamal Khuri-Makdisi from whom I learned a lot.

It would be hard for me to express how grateful I am to Christelle, the love of my life, for her unwavering support throughout this adventure. I am constantly dazzled by her brilliance, and I will never forget her heroic feats when we were separated for two long years by 5400 miles and 7 time zones.

I have been blessed with a great family whose size makes it impossible to properly acknowledge here. Let me just mention a few highly distinguished members. My grandparents Saadallah and Marie, who were particularly close to me as I was growing up, are an example of hard work and wisdom. My (so-called) uncle Charbel has been a constant source of help and support despite the fact that we have not lived in the same country for the past 15 years.

I have missed a lot the intelligence and sharp sense of humor of my sisters Maria and Elena and my brother Rabih. Their loving support has been very precious. My biggest regret is not having spent enough time with them during the last years.

I owe a lot to my parents, Bassam and Hilda. My father’s limitless generosity and support has made my life much easier. My mother instilled in me an appetite for knowledge from an early age and her unmatched dedication and love inspire me every day.
A Note on The Content

This thesis contains a subset of the author’s work during his Ph.D. [GJ18, GKR17, GS17, GHKS17, FG17, BGH+16, GKS16b, GGG16, GKS16, GKS16a, GGG15, GL15, BBG14, SAH+13, GHI+13].

Chapter 2 and Chapter 4 are based on the following previous publication:


Chapter 3 is based on the following previous publication:


Chapter 5 is based on the following two previous publications:


Chapter 6 is based on the following previous publication:


Chapter 7 is based on the following manuscript:


This research was supported in part by an NSF STC Award CCF 0939370, NSF award numbers CCF-1217423, CCF-1650733 and CCF-1420692, an Irwin and Joan Jacobs Presidential Fellowship and an IBM Ph.D. Fellowship.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Notation</td>
<td>66</td>
</tr>
<tr>
<td>3.2.2</td>
<td>The Non-Interactive Simulation Problem</td>
<td>66</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Reformulation of Gap-Non-Int-Sim</td>
<td>68</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Fourier Analysis and Multi-Linear Polynomials</td>
<td>70</td>
</tr>
<tr>
<td>3.2.5</td>
<td>Hypercontractivity and Moment Bounds</td>
<td>73</td>
</tr>
<tr>
<td>3.2.6</td>
<td>Maximal Correlation and Witsenhausen’s Rounding</td>
<td>74</td>
</tr>
<tr>
<td>3.2.7</td>
<td>Two-Dimensional Berry-Esseen Theorem</td>
<td>75</td>
</tr>
<tr>
<td>3.3</td>
<td>Main Technical Lemma and Overview</td>
<td>77</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Proof Overview</td>
<td>77</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Decidability of Gap-Non-Int-Sim</td>
<td>78</td>
</tr>
<tr>
<td>3.4</td>
<td>Smoothing of Strategies</td>
<td>79</td>
</tr>
<tr>
<td>3.5</td>
<td>Joint Regularity Lemma for Fourier Concentrated Functions</td>
<td>80</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Regularity Lemma for Constant-Degree Polynomials</td>
<td>81</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Joint Regularity Lemma</td>
<td>84</td>
</tr>
<tr>
<td>3.6</td>
<td>Applying Correlation Bounds for Low-Influence Functions</td>
<td>85</td>
</tr>
<tr>
<td>3.7</td>
<td>Simulating Correlated Gaussians</td>
<td>88</td>
</tr>
<tr>
<td>3.8</td>
<td>Putting it All Together!</td>
<td>90</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Generalizing to Arbitrary Binary Targets</td>
<td>92</td>
</tr>
<tr>
<td>3.9</td>
<td>Open Questions</td>
<td>93</td>
</tr>
<tr>
<td>4</td>
<td>Communication with Imperfectly Shared Randomness</td>
<td>95</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction &amp; Related Work</td>
<td>95</td>
</tr>
<tr>
<td>4.2</td>
<td>Proof of Theorem 4.1.1</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>Communication with Functional Uncertainty</td>
<td>101</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction &amp; Related Work</td>
<td>101</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Uncertain-Communication Complexity</td>
<td>103</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Results</td>
<td>105</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Prior Work</td>
<td>111</td>
</tr>
<tr>
<td>5.2</td>
<td>The Uncertain-Communication Complexity Model</td>
<td>112</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Communication Complexity</td>
<td>112</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Uncertain-Communication Complexity</td>
<td>113</td>
</tr>
<tr>
<td>5.3</td>
<td>Hardness of Contextual Agreement</td>
<td>115</td>
</tr>
<tr>
<td>5.4</td>
<td>One-Way Uncertain Communication: Public-Coin Protocol</td>
<td>120</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Overview of Protocol</td>
<td>120</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Rejection Sampling</td>
<td>122</td>
</tr>
<tr>
<td>5.4.3</td>
<td>Proof of Theorem 5.1.6</td>
<td>123</td>
</tr>
<tr>
<td>5.5</td>
<td>One-Way Uncertain Communication: Imperfectly Shared Randomness Protocol</td>
<td>126</td>
</tr>
<tr>
<td>5.6</td>
<td>Lower Bound on Public-Coin Protocols</td>
<td>128</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Proof of Lemma 5.6.2</td>
<td>129</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Proof of Lemma 5.6.3</td>
<td>130</td>
</tr>
<tr>
<td>5.7</td>
<td>Improved Lower Bound on One-Way Public-Coin Uncertain Protocols</td>
<td>135</td>
</tr>
<tr>
<td>5.7.1</td>
<td>Overview of Proof of Theorem 5.1.4</td>
<td>136</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.7.2</td>
<td>Proof of Theorem 5.1.4</td>
<td>138</td>
</tr>
<tr>
<td>5.7.3</td>
<td>Proof Preliminaries for Sections 5.7.4 and 5.7.5</td>
<td>140</td>
</tr>
<tr>
<td>5.7.4</td>
<td>Simulation Protocol</td>
<td>142</td>
</tr>
<tr>
<td>5.7.5</td>
<td>Proof of Lemma 5.7.3</td>
<td>147</td>
</tr>
<tr>
<td>5.8</td>
<td>Lower Bound on Private-Coin Uncertain Protocols</td>
<td>148</td>
</tr>
<tr>
<td>5.8.1</td>
<td>Overview of Proof of Theorem 5.1.11</td>
<td>148</td>
</tr>
<tr>
<td>5.8.2</td>
<td>Proof of Theorem 5.1.11</td>
<td>150</td>
</tr>
<tr>
<td>5.8.3</td>
<td>Proof of Lemma 5.8.4</td>
<td>152</td>
</tr>
<tr>
<td>5.8.4</td>
<td>Proof of Part (iii) of Theorem 5.1.11</td>
<td>153</td>
</tr>
<tr>
<td>5.9</td>
<td>Two-Dimensional Berry-Esseen Theorem for Independent Random Variables</td>
<td>158</td>
</tr>
<tr>
<td>5.10</td>
<td>The Need to Work with Positive-Error Uncertain Protocols</td>
<td>159</td>
</tr>
<tr>
<td>5.11</td>
<td>Discussion and Future Directions</td>
<td>160</td>
</tr>
<tr>
<td>5.8</td>
<td>Lower Bound on Private-Coin Uncertain Protocols</td>
<td>148</td>
</tr>
<tr>
<td>5.8.1</td>
<td>Overview of Proof of Theorem 5.1.11</td>
<td>148</td>
</tr>
<tr>
<td>5.8.2</td>
<td>Proof of Theorem 5.1.11</td>
<td>150</td>
</tr>
<tr>
<td>5.8.3</td>
<td>Proof of Lemma 5.8.4</td>
<td>152</td>
</tr>
<tr>
<td>5.8.4</td>
<td>Proof of Part (iii) of Theorem 5.1.11</td>
<td>153</td>
</tr>
<tr>
<td>5.9</td>
<td>Two-Dimensional Berry-Esseen Theorem for Independent Random Variables</td>
<td>158</td>
</tr>
<tr>
<td>5.10</td>
<td>The Need to Work with Positive-Error Uncertain Protocols</td>
<td>159</td>
</tr>
<tr>
<td>5.11</td>
<td>Discussion and Future Directions</td>
<td>160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6</th>
<th>Uncertain Distributed Compression</th>
<th>163</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction &amp; Related Work</td>
<td>163</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Model</td>
<td>163</td>
</tr>
<tr>
<td>6.1.2</td>
<td>A Motivation: Language Formation and Development</td>
<td>164</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Context and Main Benchmarks</td>
<td>165</td>
</tr>
<tr>
<td>6.1.4</td>
<td>Results</td>
<td>167</td>
</tr>
<tr>
<td>6.1.5</td>
<td>Previous Work on Language Evolution</td>
<td>168</td>
</tr>
<tr>
<td>6.2</td>
<td>Formal Definitions</td>
<td>168</td>
</tr>
<tr>
<td>6.3</td>
<td>A Distributed Compression Protocol</td>
<td>169</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Overview of the Protocol</td>
<td>170</td>
</tr>
<tr>
<td>6.3.2</td>
<td>The STATIC Protocol</td>
<td>170</td>
</tr>
<tr>
<td>6.3.3</td>
<td>The UNCERTAIN Protocol</td>
<td>170</td>
</tr>
<tr>
<td>6.3.4</td>
<td>Final Protocol</td>
<td>174</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Efficient Implementation</td>
<td>179</td>
</tr>
<tr>
<td>6.4</td>
<td>A Computationally Inefficient Protocol with Smaller Communication</td>
<td>181</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Analysis of Computationally Inefficient Protocol</td>
<td>183</td>
</tr>
<tr>
<td>6.5</td>
<td>Conclusion &amp; Future Directions</td>
<td>186</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7</th>
<th>Correlated Sampling</th>
<th>189</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introduction &amp; Related Work</td>
<td>189</td>
</tr>
<tr>
<td>7.1.1</td>
<td>Special Cases</td>
<td>191</td>
</tr>
<tr>
<td>7.1.2</td>
<td>Open Questions and Future Work</td>
<td>192</td>
</tr>
<tr>
<td>7.2</td>
<td>Lower Bounds on Correlated Sampling</td>
<td>193</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Lower Bound in a Special Case</td>
<td>196</td>
</tr>
<tr>
<td>7.3</td>
<td>Correlated Sampling Strategies</td>
<td>197</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Strategy in a Special Case</td>
<td>198</td>
</tr>
</tbody>
</table>

| 8       | Conclusion                                                           | 201  |
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Uncertainty in Communication</td>
<td>16</td>
</tr>
<tr>
<td>2-1</td>
<td>Common Randomness Generation. Here, $K_A$ denotes Alice’s output, $K_B$</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>denotes Bob’s output, and $C$ denotes then number of bits of communication.</td>
<td></td>
</tr>
<tr>
<td>3-1</td>
<td>Example Source Distribution for Which Many Samples Are Needed.</td>
<td>65</td>
</tr>
<tr>
<td>3-2</td>
<td>Non-Interactive Simulation of Joint Distributions [KA12, KA15]</td>
<td>67</td>
</tr>
<tr>
<td>3-3</td>
<td>Dependency of Parameters in the Proof of Theorem 3.3.1</td>
<td>91</td>
</tr>
<tr>
<td>4-1</td>
<td>Simultaneous Message Passing Communication Model</td>
<td>96</td>
</tr>
<tr>
<td>5-1</td>
<td>Communication with Functional Uncertainty</td>
<td>105</td>
</tr>
<tr>
<td>5-2</td>
<td>Contextual Agreement Problem</td>
<td>116</td>
</tr>
<tr>
<td>5-3</td>
<td>Operation of $\text{Stretch}_{r,a}$ Procedure for $r = 2$ and $a = 3$.</td>
<td>155</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction: Uncertainty in Communication

In many forms of communication, two or more agents interact with each other in order to achieve a certain goal such as transmitting and receiving information, computing a function of their input datasets or agreeing on an desirable object (such as a random ID or a secret key). A fundamental challenge in communication is raised by various forms of uncertainty, which can arise as noise in the inputs or in the exchanged messages, as mismatches in the shared randomness used by the distributed agents to speed up communication, or as ambiguity in the shared context or goal of the communication.

This thesis studies these diverse forms of uncertainty, building on work done on the topic in several disciplines, including information theory, computational complexity and distributed computing. In some cases, we answer open questions raised by previous work in the field; in other cases, we study novel generalized models of communication capturing forms of uncertainty that had eluded previous widely studied models.

Figure 1-1 illustrates the various components of uncertain communication that we elaborate on in the next sections and study throughout this thesis.

While error correction is a classical paradigm for coping with the uncertainty introduced by a noisy channel, uncertainty can also occur in several other forms if the communicating players do not completely know (or understand) each other. We refer to this as contextual uncertainty. Specifically, in many forms of communication, the interacting parties share a context that is i) huge and ii) imperfectly shared. Examples of this phenomenon range from human communication where people with extremely different backgrounds and experiences are still able to communicate very efficiently, to the emerging area of “conversational artificial intelligence” which underlies intelligent virtual assistants, and to mathematical proofs that are not written in any formal system of logic but are yet considered correct. What makes such efficient forms of communication possible? How can we mathematically model uncertainty in context and approach such problems?

In the following sections, we present our results within this area.
1.1 Uncertainty in Shared Randomness

Randomness typically plays an important role in speeding up centralized algorithms as well as in decreasing the communication cost and time complexity of distributed protocols. Ideally, the different players in a distributed protocol would have joint access to an infinitely long string of random bits. Although it leads to significant speed-ups in numerous setups, such a perfect sharing of randomness might be practically infeasible. Realistically, it is more likely for the players to have access to weaker forms of correlation (such as correlated binary sequences or correlated Gaussian signals) than to random bits in perfect agreement. Can imperfect correlations be used to agree on a random id or on a secret key? Can they allow significant gains in communication as does perfectly shared randomness? We obtain several improvements on the state-of-the-art for these questions (answering some open problems from previous work) which we describe next in more detail.

1.1.1 Common Randomness & Secret Key Generation

In the setup of common randomness generation, two players are given independent identically distributed samples from a known distribution, and are required to agree on a given number of random bits while minimizing the communication and maximizing the agreement probability. This question has been studied in information theory, theoretical computer science and cryptography where it is equivalent to the secret key generation problem in which case the generated random sequence of bits has to be unknown to any eavesdropper (e.g., [AC93, AC98]). This setup has several important applications. For instance, the generated random key can be used as shared randomness by distributed protocols, and the generated secret key can be used for encryption. This setting is in fact closely related to the identification capacity [AD89] and to hardware-based procedures for extracting a unique random ID from
process variations [LLG+05, SHO08, YLH+09] that can be used in authentication [LLG+05, SD07].

**Previous Work on Common Randomness & Secret Key Generation** In information theory, common randomness and secret key generation have been extensively studied since the seminal work of Ahlswede and Csiszar [AC93, AC98] who defined (an amortized version of) the setup and obtained several results on special cases of the problem. In particular, in the case of one-way communication, they characterized (amortized) common randomness generation for an arbitrary source of correlation in terms of its strong data processing constant, a beautiful mathematical measure that turns out to have important applications in information theory, and more recently in theoretical computer science.

Since the works of Ahlswede and Csiszar, numerous follow-up works studied common randomness and secret key generation including the recent work of Liu, Cuff and Verdu [LCV15, LCV16] who in particular gave a characterization of $r$-round common randomness generation in terms of generalization of the strong data processing constant.

In theoretical computer science, Bogdanov and Mossel [BM11] studied common randomness generation in the zero-communication case and when the source consists of $\rho$-correlated bits, and they gave a protocol with the optimal trade-off between agreement probability, the number of generated random bits and the correlation parameter $\rho$. Guruswami and Radhakrishnan [GR16] generalized the results of Bogdanov and Mossel to one-way communication, in which case they obtained the tight trade-off between communication, agreement probability, number of generated random bits and $\rho$. While the protocols of [BM11] and [GR16] achieve the optimal trade-offs, they are non-explicit, sample-inefficient and computationally inefficient. Obtaining more efficient protocols is an open question that was explicitly asked by [BM11].

**Our Results on Common Randomness & Secret Key Generation** In the two-party zero-communication and one-way communication setups, we show in [GJ18] that certain explicit families of error-correcting codes (dual-BCH codes and their variants in Euclidean space) can be used to significantly improve the sample efficiency of common randomness generation, answering an open question of [BM11]. Specifically, we consider arguably the two most natural sources of correlation: $\rho$-correlated bits and $\rho$-correlated Gaussians. For each of these two sources, we give schemes for generating $k$ random bits using $n = \text{poly}(k)$ samples and achieving the optimal dependence of agreement probability on $k$ and $\rho$ in the zero-communication case, and the optimal trade-off between communication and agreement probability in terms of $k$ and $\rho$ in the one-way communication case. We point out that while our schemes are explicit and sample-efficient, there are not computationally efficient. Obtaining computationally efficient schemes remains a fascinating open question.

We also reveal a surprising novel connection between interactive (amortized) common randomness generation and information complexity [CSYW01, BYJKS02, Bra15]. To describe this connection, recall that the external information cost of a protocol is the amount of information it reveals about the inputs to an external observer, while its internal informa-
tion cost is the amount of information it reveals to each of the two players about the other player’s input. Surprisingly, we show for any source of correlation and for any number of rounds, the largest achievable ratio of generated entropy to communication is equal to the largest achievable ratio of external to internal information costs.

The full details can be found in Chapter 2.

1.1.2 Non-Interactive Simulation of Joint Distributions

A common assumption in the design of distributed algorithms is to allow the different parties access to a string of shared randomness in the form of independent random bits. However, many natural sources of correlation do not come in the form of bits that are perfectly (i.e., identically) shared by the different parties. Instead, the parties can often have access to correlated but non-identical random variables. Is it possible to convert one form of correlation to a more useful one? More generally, if Alice and Bob are given independent and identically distributed samples from a distribution $P$, can they generate a sample from another distribution $Q$ without interaction? This question – which dates back to classical work in information theory [GK73, Wyn75] and which has been studied by several subsequent works (see, e.g., [KA15] and the references within) – belongs to the class of problems on tensor powers on graphs [Alo02]. Understanding the computational aspects of such questions are notorious open problems: they are not known to be decidable while not even known to be NP-hard!

Our Result on Non-Interactive Simulation We prove [GKS16b] that the above non-interactive simulation problem is decidable for binary output alphabets. To do so, we significantly leveraged recent developments in theoretical computer science, and in particular, the invariance principle from the analysis of Boolean functions [MO05]. Our work allows different parties to convert very messy sources of correlation to a very structured one: correlated marginally-uniform random bits.

Future Developments In a follow-up work, De, Mossel and Neeman showed the decidability of the non-interactive simulation problem for arbitrary discrete output alphabets (including the particular case of computing the noise stability of a function) [DMN17, DMN18]. In a recent work (that is not included in this thesis) [GKR17], we give a dimension reduction for polynomials over the Gaussian space which we use to significantly improve the bound on the running time of the decidability procedure of [DMN17, DMN18] for non-interactive simulation (and for the particular case of computing the noise stability). Our dimension reduction can be seen as a generalization of the Johnson-Lindenstrauss lemma (which has been extremely influential in computer science with numerous applications including unsupervised learning, compressed sensing, manifold learning, and graph embedding).

The full details can be found in Chapter 3.
1.1.3 Communication with Imperfectly Shared Randomness

Communication complexity (first introduced by Yao [Yao79]) studies the amount of communication needed by two or more parties in order to compute a joint function of their inputs. For many important functions, such as testing equality or closeness of two input datasets, there are well-known powerful protocols that are able to compute the function with very little communication assuming the players have access to perfectly shared randomness (i.e., identical sequences of random bits). What if the parties share the randomness imperfectly (i.e., they are given correlated random strings that are not necessarily equal)? The parties can choose to ignore their imperfectly shared randomness and only use their private randomness; this however can lead to significant overhead in communication. The main question is whether the parties can use the imperfectly shared randomness in order to obtain much more efficient protocols.

Previous Work on Communication with Imperfectly Shared Randomness This setup was first independently studied by Bavarian, Gavinsky and Ito [BGI14] in the simultaneous message passing model (where two players can each send a single message to a referee who should then output the answer), and by Canonne, Guruswami, Meka and Sudan [CGMS14] in the case of one-way and interactive communication. In both of these frameworks, Alice and Bob wish to compute a joint function of their inputs and have access to i.i.d samples from a known source. For the sake of exposition, we here consider the most natural source of correlation: correlated bits (although our result holds for more general sources).

In the one-way and interactive settings, [CGMS14] showed that any function having a two-way protocol using perfectly shared randomness and with communication $k$ bits, has a one-way protocol with imperfectly shared randomness and with communication at most $2^{O(k)}$ bits. On the negative side, they showed the existence of a function for which the one-way communication with perfectly shared randomness is equal to $k$ bits whereas the two-way communication with imperfectly shared randomness is at least $2^{Ω(k)}$ bits.

In the simultaneous message passing setting, [BGI14] showed that there exists a function whose communication with perfectly shared randomness is equal to $k$ bits, but for which any protocol with imperfectly shared randomness has communication at least $2^{Ω(k)}$ bits.

Our Result on Communication with Imperfectly Shared Randomness The previous work of [CGMS14] and [BGI14] left the following piece missing from the picture: does any function having a simultaneous message passing protocol with perfectly shared randomness and communication $k$ bits, necessarily have a simultaneous message passing protocol with imperfectly shared randomness and communication $2^{O(k)}$ bits? In [GJ18], we show that this is indeed the case. Building on the approach of [CGMS14], our solution gives a simple optimal protocol for estimating $ℓ_2$-distances using imperfectly shared randomness in the simultaneous message passing model.

The full details can be found in Chapter 4.
1.2 Uncertainty in Goal of Communication

Functional Uncertainty  One type of contextual uncertainty is the one pertaining to the goal of the communication. Namely, how can we capture the setup where two communicating parties have somewhat similar views regarding the goal of the communication without requiring these views to be perfectly aligned?

Quite often, the goal of the communication is to compute a function of the parties’ inputs. In this case, the uncertainty in the goal of the communication takes the form of functional uncertainty, which we study.

Our Results on Communication with Functional Uncertainty  In [GKKS16], we suggest a framework for studying functional uncertainty by building on the model of communication complexity introduced by Yao [Yao79]. In this framework, Alice thinks that the goal of the communication is to compute function $f$, Bob thinks that the goal is to compute function $g$, where $f$ and $g$ are close but not necessarily identical. The particular case where $f = g$ corresponds to the widely studied communication complexity model. On the other hand, when $f \neq g$, Alice and Bob can choose to ignore part of their context by exchanging their entire inputs; this would allow Alice to compute $f$ and Bob to compute $g$ albeit with an enormous communication cost. The main question is therefore whether Alice and Bob can use their (imperfectly) shared context in order to solve the problem much more efficiently.

We answer this question by designing a (randomized) protocol that allows a sender and a receiver to overcome their uncertainty about the goal of the communication with a relatively small overhead (i.e., with a small blow-up in communication compared to the case where $f = g$). In [GKKS16] and in a follow-up work [GS17], we study the power of public and private-coin protocols as well as protocols with imperfectly shared randomness, and we show how results from communication complexity and graph theory can be leveraged in order to better understand the randomness requirements of efficient protocols for communication with functional uncertainty.

The full details can be found in Chapter 5.

1.3 Uncertainty in Prior Distribution

1.3.1 Uncertain Distributed Compression  We now turn to a different type of contextual uncertainty, where the parties are unsure about the inputs’ prior distribution. In the special case where the players know the prior distribution, the problem is an instance of classical data compression where known solutions (such as Huffman codes) apply. On the other hand, when the players are uncertain about the prior distribution, they can choose to ignore any information they have about it and use a canonical encoding of the universe elements into bits (regardless of their probabilities); this trivial scheme would solve the problem but with huge communication. Thus, the main
question is whether one can use the players' uncertain knowledge of the prior distribution in order to come up with much more efficient compression algorithms.

Our Results: Model and Protocol

**Uncertain Distributed Compression: The Model** Namely, we study the setup where $K$ players are trying to learn an unknown distribution based on individual samples and have to agree on a compression/decompression scheme that should compress this unknown distribution well despite disparities in their samples. In our model, a sequence of pairs of players from a set of $K$ players are chosen and tasked to communicate messages drawn from an unknown distribution $Q$. The only knowledge that the players have about $Q$ is from previously drawn samples. Since these samples differ from player to player, at any point of time different players will have different priors about the message distribution (and these priors are all likely to be different from the true distribution). The only common knowledge between the players is restricted to a common prior distribution $P$ (which can be quite far from the true distribution $Q$) and some constant number of bits of information (such as a learning algorithm).

Classical solutions require all players to agree on the “same” approximation $Q'$ to $Q$ and then achieve an expected compression length of $H(Q) + D(Q||Q')$ bits, where $D(\cdot||\cdot)$ denotes the KL-divergence between distributions. Players can settle for $Q' = P$ (so not learn at all) and achieve weak compression or try for something better by communicating a $Q'$ that someone has learned. However, communicating any reasonably accurate distribution $Q'$ leads to enormous communication.

**Uncertain Distributed Compression: Efficient Protocol** In [GHKS17], we present a novel decentralized solution with significantly better performance: We give a natural, uniform and efficient algorithm that compresses the communication down to an average cost per message of $O(H(Q) + \log(D(Q||P))$ bits in $\Theta(K)$ iterations. Our algorithm builds on recent work on uncertain compression in the two-party setup by [JKKS11] and [HS14a].

The full details can be found in Chapter 6.

1.3.2 Correlated Sampling

Whenever two parties have knowledge of two distributions that are not necessarily equal (as in the compression setup described in the previous section), one basic question is for the parties to estimate the distance between these two distributions. One of the most widely used notion of distance between distributions is the total variation distance. Correlated Sampling is an algorithmic question that can be used to estimate the total variation distance between two distributions held by two parties.

Formally, a correlated sampling protocol is one where Alice is given a distribution $P$, Bob is given a distribution $Q$ over the same universe and their goal is to non-interactively
sample from their respective distributions while minimizing the disagreement probability. This very basic algorithmic procedure is used in several areas including sketching algorithms, approximation algorithms based on rounding linear programming relaxations, the study of parallel repetition and very recently cryptography. A well-known protocol – variants of which are due to Broder [Bro97], Kleinberg and Tardos [KT02], and Holenstein [Hol07] – solves this question while achieving disagreement probability at most $\frac{2\delta}{1 + \delta}$ where $\delta$ is the total variation distance between $P$ and $Q$. This protocol has been rediscovered numerous times across different communities.

**Our Result on Correlated Sampling** In [BGH+16], we give a surprisingly simple proof that this protocol is in fact tight! Namely, we show that any protocol solving the correlated sampling problem on distributions at total variation distance $\delta$ cannot have disagreement probability smaller than $\frac{2\delta}{1 + \delta}$.

The full details can be found in Chapter 7.

### 1.4 Thesis Roadmap

In Chapter 2, we give our results on common randomness generation. Our results on the non-interactive simulation of joint distributions appear in Chapter 3. Our result on correlated sampling are given in Chapter 7. Our protocol for communication with imperfectly shared randomness in the simultaneous message passing model is given in Chapter 4. The description of our communication with functional uncertainty model and our corresponding protocols and lower bounds appear in Chapter 5. Our results on uncertain distributed compression are given in Chapter 6. Finally, we conclude with some future directions in Chapter 8.
Chapter 2

Common Randomness and Secret Key Generation

2.1 Introduction & Related Work

In this chapter, we present our results on common randomness and secret key generation.

Common randomness plays a fundamental role in various problems within cryptography and information theory. We study this problem (depicted in Figure 2-1) in the basic two-party communication setting in which Alice and Bob wish to agree on a (random) key by drawing i.i.d. samples from a known source such as correlated bits or correlated Gaussians. If we further require that an eavesdropper, upon seeing the communication only, gains no information about the shared key, then this defines a secret key scheme. This information-theoretic approach to security was introduced in the seminal works of Mauer [Mau93] and Ahlswede and Csiszár [AC93]. Both common randomness and secret-key generation have been extensively studied in information theory [AC98, CN00, GK73, Wyn75, CN04, ZC11, Tya13, LCV15, LCV16]. Common randomness has applications to identification capacity [AD89] and hardware-based procedures for extracting a unique random ID from process variations [LLG+05, SHO08, YLH+09] that can be used in authentication [LLG+05, SD07].

Randomness is a powerful tool as well in the algorithm designer’s arsenal. Shared keys (aka public randomness) are used crucially in the design of efficient communication protocols with immediate applications to diverse problems in streaming, sketching, data structures and property testing. Common randomness is thus a natural model for studying how shared keys can be generated in settings where it is not available directly [MO04, MOR+06, BM11, CMN14, CGMS14, GR16]. In this chapter, we take the approach of treating correlated sources as a critical algorithmic resource, and ask whether common randomness can be generated efficiently.\footnote{1Notably, the schemes that we design can also be easily transformed into secret key schemes, as shown later.}

For $-1 \leq \rho \leq 1$, we say that $(X, Y) \sim \text{DSBS}(\rho)$ (doubly symmetric binary source with correlation parameter $\rho$) if $X, Y$ are both uniform over $\{\pm 1\}$ and their correlation (and
Figure 2-1: Common Randomness Generation. Here, $K_A$ denotes Alice’s output, $K_B$ denotes Bob’s output, and $C$ denotes the number of bits of communication.

covariance) is $\mathbb{E}[XY] = \rho$ (i.e., a binary symmetric channel with uniform input). We say that $(X, Y) \sim \text{BGS}(\rho)$ (bivariate Gaussian source with correlation parameter $\rho$) if $X, Y \sim \mathcal{N}(0, 1)$, the standard normal distribution, and their correlation is again equal to $\rho$.\(^2\)

Bogdanov and Mossel [BM11] gave a common randomness scheme for DSBS$(\rho)$ with zero-communication to generate $k$-bit keys that agree with probability $2^{-\frac{1+\rho}{1+\rho}k}$, up to lower order inverse poly$(k, 1 - \rho)$ factors (which we suppress henceforth). Using the hypercontractive properties of the noise operator [Bon70, Bec75], they also proved the “converse” result, i.e., that this bound on the agreement (probability) is essentially the best possible. In a follow-up work, Guruswami and Radhakrishnan [GR16] recently gave a one-way scheme that achieves an optimal trade-off between communication and agreement probability.\(^3\) Note that a simple scheme in which Alice just sends her input requires $k - O_\rho(1)$ bits of communication for constant agreement. In contrast, their scheme can guarantee the same agreement using only $(1 - \rho^2) \cdot k$ bits of communication. This is a non-trivial bound since for $\rho > 0$, the ratio of entropy to communication ($= 1/(1 - \rho^2)$) is strictly bounded away from 1 as $k \to \infty$. On the other hand, the above schemes are non-explicit (they are proved using the probabilistic method) and use an exponential number of samples in $k$. Bogdanov and Mossel [BM11] asked whether an explicit and efficient scheme can be designed, motivating the definition below.

\(^2\)Note that BGS$(\rho)$ is uniquely defined. To see this, note that any linear combination of $X$ and $Y$ is also Gaussian; thus, $(X, Y)$ has a multivariate Gaussian distribution and is hence characterized by its first and second moments.

\(^3\)They also use hypercontractivity to prove the converse, which extends to other sources including BGS$(\rho)$. 24
We say that a common randomness scheme to generate $k$-bit keys (with $k$ as input) is resource-efficient, if it (i) is explicitly defined, (ii) uses $\text{poly}(k)$ samples, (iii) has constant agreement probability, and (iv) achieves an amortized ratio of entropy to communication bounded away from 1. We give the first efficient scheme for correlated bits and Gaussians, answering the question of [BM11]:

**Theorem 2.1.1.** There exist resource-efficient one-way common randomness schemes for $\text{DSBS}(\rho)$ and $\text{BGS}(\rho)$ using $(1 - \rho^2) \cdot k$ bits of communication. For zero-communication, there exist explicit schemes for $\text{DSBS}(\rho)$ and $\text{BGS}(\rho)$ using $\text{poly}(k)$ samples with agreement probability $2^{-\frac{1-\rho^2}{1+\rho} \cdot k}$, up to polynomial factors.

More generally, we obtain explicit one-way schemes with optimal trade-off between communication and agreement probability, matching [GR16], while using only $\text{poly}(k)$ samples. Below is the formal statement.

**Theorem 2.1.2.** Let $0 < \rho < 1$ and $0 \leq \delta \leq \sqrt{1 - \rho^2}$ be arbitrary. Set $\varphi = \rho + \delta \sqrt{1 - \rho^2}$. Then, there exist explicit one-way common randomness schemes for $\text{DSBS}(\rho)$ and $\text{BGS}(\rho)$ using $\text{poly}(k)$ samples such that:

1. the entropy of the key is at least $k - o(k)$;\footnote{By an explicit scheme, we mean that its existence is not proved using the probabilistic method, and that the scheme can be constructed in time $\text{poly}(k, n)$ where $k$ is the number of generated common random bits and $n$ is the number of samples drawn from the source.}

2. the agreement probability is at least $2^{-\varphi^2 k}$, up to polynomial factors; and

3. the communication is $O((1 - \varphi^2) \cdot k)$ bits.

We point out that our schemes are resource efficient but computationally inefficient. One representative challenge that arises here is in decoding dual-BCH codes, which are an explicit algebraic family of error-correcting codes, from a very large number of errors.

The above schemes follow a template that generalizes the approach taken by [BM11, GR16]. It relies on a carefully constructed codebook $\mathcal{C} \subseteq \mathbb{R}^n$ of size $2^k$, where $n$ is the number of samples. Alice outputs the codeword in $\mathcal{C}$ with the largest projection while Bob does the same on a subcode of $\mathcal{C}$ based on Alice’s message. The analysis of the template reduces it to the problem of obtaining good tail bounds on the joint distribution induced by these projections. For $\text{BGS}(\rho)$, we use a codebook consisting of an explicitly defined large family of nearly-orthogonal vectors in $\mathbb{R}^n$ due to Tao [Tao13], who showed their near-orthogonality property using the Weil bound for curves. The novel part of the analysis involves getting precise conditional probability tail bounds on trivariate Gaussians induced by the projections, whose covariance matrix has a special structure. Standard methods only give asymptotic bounds on such tails which are inadequate in the low-communication regime. Here, the best possible agreement is exponentially small in $k$. Our analysis determines the exact constant in the exponent by carefully evaluating the underlying triple integrals.

\footnote{We follow [GR16] who actually consider the min-entropy of Alice’s output, which is justifiable on technical grounds.}
The resource-efficient scheme for DSBS(\(\rho\)) is based on dual-BCH codes that can be seen as an \(\mathbb{F}_2\)-analogue of Tao’s construction. The Weil bound for curves implies that dual-BCH codes are “unbiased”, in the sense that any two distinct codewords are at distance \(\approx n/2\) (with \(n\) being the block length). Analogous to the Gaussian case, the analysis involves getting precise bounds on the (conditional) tail probabilities of various correlated binomial sums. Since \(n = \text{poly}(k)\), we cannot handle these binomial sums using the (two-dimensional) Berry-Esseen theorem, since the incurred additive error of \(1/\sqrt{n}\) would overwhelm the target agreement probability. Moreover, crude concentration and anti-concentration bounds cannot be used since they do not determine the exact constant in the exponent. We directly handle these correlated binomial sums, which turns out to involve some tedious calculations related to the binary entropy function.

Interactive Common Randomness and Information Complexity. Ahlswede and Csiszár [AC93, AC98] studied common randomness in their seminal work using an amortized communication model. They defined it as the maximum achievable ratio \(a/c\), such that for every large enough number of samples \(n\), Alice and Bob can agree on a key of \(a \cdot n\) bits using \(c \cdot n\) bits of communication, where the agreement probability tends to 1 (as \(n\) tends to infinity). This more stringent linear relationship between the quantities is not obeyed by our explicit schemes. For one-way communication, they characterized this ratio in terms of the Strong Data Processing Constant of the source, which is intimately related to its hypercontractive properties [AG76, AGKN13]. More recently, Liu, Cuff and Verdu [LCV15, LCV16, Liu16] extended this beyond one-way communication. In particular, [Liu16] derives the “rate region” for \(r\)-round amortized common randomness.

In this chapter, we show that \(r\)-round amortized common randomness can be alternatively characterized in terms of two well-studied notions in theoretical computer science: the internal and external information costs of communication protocols. Recall that the internal information cost [BJKS04, BBCR13] of a two-party randomized communication protocol is the total amount of information that each of the two players learns about the other player’s input, whereas its external information cost [CSWY01] is the amount of information that an external observer learns about the inputs (see section 2.5 for formal definitions). These measures have been extensively studied within the context of communication complexity. While being interesting measures in their own rights, they have also been the central tool in tackling direct-sum problems, with numerous applications, e.g., in data streams and distributed computation.

**Theorem 2.1.3** (Informal Statement). *Given an arbitrary distribution \(\mu\), let \(\Gamma_r\) denote the supremum over all \(r\)-round randomized communication protocols \(P\) of the ratio of the external information cost to the internal information cost of \(P\) with respect to \(\mu\). Then, for \(r\)-round amortized common randomness, \(\Gamma_r\) equals the largest achievable ratio \(H/R\) such that using \(\mu\) as the source, for every large enough \(n\), Alice and Bob can agree on a key of \(H \cdot n - O(1)\) bits with probability \(1 - o_n(1)\) using \(r\) rounds and \(R \cdot n + O(1)\) bits of communication.*

---

6For more on unbiased codes, we refer the reader to the work of Kopparty and Saraf [KS13].
For the proof, we use a direct-sum approach, a classical staple of information complexity arguments. Our setup is slightly different from the known direct-sum results because we need to lower bound the internal information cost of the \(n\)-input protocol as well as upper bound its external information cost (which is non-standard) simultaneously. The essential ingredient is the same: embed the input on a judiciously chosen coordinate, but the argument works on a round-by-round basis so as to keep the mutual information expressions intact. To prove the other direction, we use the rate region of [LCV16, Liu16] to get a lower bound on \(\Gamma_r\).

We now outline our results in various settings where common randomness plays an important role.

**Secret Key Generation:** While secret key generation requires common randomness, in the amortized setting they are known to imply each other [LCV16, Liu16]: the rate pair \((H, R)\), using the notation of theorem 2.1.3, is achievable for common randomness if and only if \((H - R, R)\) is achievable for secret key generation. In particular, using the Strong Data Processing Constant for DSBS(\(\rho\)), the rate ratio \(H/R = 1/(1 - \rho^2)\) is achievable for common randomness and the rate ratio \(\rho^2/(1 - \rho^2)\) for secret key generation, but using non-explicit schemes. Our resource-efficient but non-amortized schemes given in theorem 2.1.1 can be easily transformed into secret key schemes. See remark 2.3.2.

**General Sources:** Theorem 2.1.2 also implies an explicit scheme for an arbitrary source \(\mu\) in terms of its maximal correlation \(\rho(\mu)\) [Hir35, Geb41, Rén59]. For \((X, Y) \sim \mu\), recall that \(\rho(\mu) := \sup \mathbb{E}F(X)G(Y)\) over all real-valued functions \(F\) and \(G\) with \(\mathbb{E}F(X) = \mathbb{E}G(Y) = 0\) and \(\text{Var } F(X) = \text{Var } G(Y) = 1\). This uses the idea (implicit in [Wit75]) that given i.i.d. samples from any source of maximal correlation \(\rho\), there is an explicit strategy (using the bivariate Central Limit Theorem) that allows Alice and Bob to use these samples in order to generate standard \(\rho\)-correlated Gaussians. The resulting scheme however is not resource-efficient.

**Correlated Randomness Generation:** In this relaxation proposed by [CGMS14], Alice and Bob are given access to DSBS(\(\rho\)) and wish to generate \(k\) bits that are jointly distributed i.i.d. according to DSBS(\(\rho'\)) where \(\rho' > \rho\). Note that the \(\rho' = 1\) corresponds to the the common randomness setup studied above. We partially answer a question of [CGMS14] by showing that even a modest improvement in the correlation requires substantial communication. Let \(\epsilon^' \log(1/\epsilon') \ll \epsilon < \frac{1}{2}\) be fixed. We show that for Alice and Bob to produce \(k\) samples according to DSBS\((1 - 2\epsilon')\) using DSBS\((1 - 2\epsilon)\) as the source requires \(\Omega(\epsilon \cdot k)\) bits of communication (even for interactive protocols and even when the agreement probability is as small as \(2^{-o(k)}\)). See section 2.8 for a detailed description.

**Locality Sensitive Hashing (LSH):** A surprising “universality” feature of our schemes (as well as previous ones) for DSBS(\(\rho\)) and BGS(\(\rho\)) using zero communication is that their definition is oblivious to \(\rho\); only the analysis for every fixed \(\rho\) shows that they have near-optimal agreement. This has a close resemblance to schemes used in LSH. Indeed, we show that our common randomness scheme leads to an improvement in the “\(\bar{\rho}\)-parameter” [IM98]
that governs one aspect of the performance of an LSH scheme. While this is mathematically interesting, we caution the reader that this does not lead to better nearest-neighbor data structures since the improvement is only qualitatively better and our scheme is computationally inefficient. See section 2.9 for more details. As discussed in that section and in Section 2.11, there seem to be a strong intuitive connection between common-randomness generation and LHS, and a very interesting question is whether one can formalize this connection.

Organization. Section 2.2 describes the template used for the one-way schemes and sets up the structure of the analysis. Section 2.3 and Section 2.4 describe the schemes for \( \text{BGS}(\rho) \) and \( \text{DSBS}(\rho) \) and their analysis. In Section 2.5, we show the connection between amortized common randomness and information complexity. In Section 2.6, we derive some properties of bivariate Gaussians that are used in Section 2.3. In Section 2.7, we derive some bounds on correlated binomial sums that are used in Section 2.4. In Section 2.8, we present our results on correlated randomness generation. In Section 2.9, we elaborate more on the intuitive connection between common randomness generation and locality sensitive hashing. In Section 2.10, we prove the min-entropy lower bounds for the schemes presented in Section 2.3 and Section 2.4. In Section 2.11, we conclude with some very intriguing open questions.

2.1.1 Preliminaries

Notation. For a tuple \( U = (U_1, U_2, \ldots, U_n) \), let \( U^j_i := (U_i, U_{i+1}, \ldots, U_j) \), when \( 1 \leq i \leq j \leq n \), and empty otherwise; we may drop the subscript when \( i = 1 \). For a distribution \( \mu \), let \( \mu^\otimes n \) be obtained by taking i.i.d. samples \( (X_1, Y_1), \ldots, (X_n, Y_n) \) from \( \mu \). Abusing notation, we say that \( (X^n, Y^n) \sim \mu^\otimes n \). Let \( \langle , \rangle \) denote the standard inner product and let \( \| \cdot \| \) denote the Euclidean norm over \( \mathbb{R} \). For any positive integer \( n \), let \( [n] := \{1, \ldots, n\} \). Let \( a \preceq b \) denote \( a \leq Cb \) for some positive global constant \( C \).

Bivariate Gaussians. Let \( (X, Y) \sim \text{BGS}(\rho) \). Let \( Q(t) := \Pr[X > t] \) denote the Gaussian tail probability and \( L(t, \varphi; \rho) := \Pr[X > t, Y > \varphi t] \) denote the (asymmetric) orthant probability. In section 2.6, we prove the following proposition, which also uses some seemingly new properties of \( Q(t) \).

Proposition 2.1.4. Let \( t, \delta \geq 0 \). Set \( \varphi := \rho + \delta \sqrt{1 - \rho^2} \) and \( \lambda_0 := \sqrt{\frac{2}{\pi}} \). Then:

\[
(a) \quad \frac{e^{-t^2/2}}{t + \lambda_0} \preceq Q(t) \preceq \frac{e^{-t^2/2}}{t + 1/\lambda_0} \leq e^{-t^2/2}; \quad \quad (b) \quad \frac{Q(t)^\delta^2}{\delta t + \lambda_0} \preceq Q(\delta t) \preceq Q(t)^\delta (t + \lambda_0)^c; \\
(c) \quad L(t, \varphi; \rho) \preceq Q(t)Q(\delta t); \quad \text{and} \quad \quad (d) \quad Q(t) \preceq Q(\delta t) \preceq Q(t)^\delta, \quad \text{if} \quad \delta \leq 1
\]

Proposition 2.1.5 (Elliptical symmetry). Let \( (X, Y) \sim \text{BGS}(\rho)^\otimes n \) and \( v, w \in \mathbb{R}^n \) have unit norm. Then, \( (\langle v, X \rangle, \langle w, Y \rangle) \sim \text{BGS}(\rho(|\langle v, w \rangle|)) \).
2.2 Template One-Way Scheme and its Analysis

The one-way schemes (including zero-communication as a special case) have the following template. Let $\mu$ denote the source on $\mathbb{R} \times \mathbb{R}$. Alice and Bob will generate $n$ i.i.d. samples from $\mu$ and use them to output $k$-bit keys. This is achieved by the players using a special codebook $C$ of $2^k$ points in $\mathbb{R}^n$ where each codeword $v \in C$ corresponds to a $k$-bit message $D(v)$, where $D : C \rightarrow \{0, 1\}^k$. For $c \geq 1$, the players also agree on a coloring $\chi$ of $C$ using $2^c$ colors such that each color class has size at most $|C| \cdot 2^{-c} + 1$. In addition, let $\diamond$ denote an auxiliary color. Thus, each color can be specified using $c + 1$ bits. For the special case of zero communication, we assume without loss of generality that all codewords are colored $\diamond$ and we set $c = 0$.

Let $t$ and $s$ be parameters that govern the achievable min-entropy and agreement probability. Let $\kappa_A$ and $\kappa_B$ be any explicit mappings such that $\kappa_A(X)$ and $\kappa_B(Y)$ are each uniformly distributed over $\{0, 1\}^k$.

### Protocol 1 One-way scheme for source $\mu$

**CR($k, \mu$):** (Goal is to generate $k$-bit common random key using source $\mu$.)

1. Let $(X, Y) \sim \mu^\otimes n$. (Alice gets $X$ and Bob gets $Y$.)

2. If $\exists$ unique $v \in C$ such that $\langle v, X \rangle > t$, Alice outputs $D(v)$ and sends $\chi(v)$. Else, Alice outputs $\kappa_A(X)$ and sends $\diamond$.

3. Bob receives the color $\tau$.

4. If $\exists$ unique $w \in C$ such that $\chi(w) = \tau$ and $\langle w, Y \rangle > s$, then Bob outputs $D(w)$. Else, Bob outputs $\kappa_B(Y)$.

The pseudocode is given in Protocol 1. For the analysis, define the following quantities:

1. **Univariate tail:** $U := \max_{v \in C} \Pr[\langle v, X \rangle > t, \langle v, Y \rangle > s]$.

2. **Bivariate tail:** $B := \min_{v \in C} \Pr[\langle v, X \rangle > t, \langle v, Y \rangle > s]$.

3. **Conditional trivariate tails:**
   (a) $T_A := \max_{v \neq w \in C} \Pr[\langle v, X \rangle > t \mid \langle v, Y \rangle > s, \langle v, Y \rangle > s]$ and (b) $T_B := \max_{v \neq w \in C} \Pr[\langle w, Y \rangle > s \mid \langle v, X \rangle > t, \langle v, Y \rangle > s]$.

**Theorem 2.2.1.** The min-entropy of the basic scheme is at least $-\log(U + 2^{-k})$. Assume that $|C| \cdot T_A \leq \frac{1}{4}$ and $|C| \cdot T_B \leq \frac{1}{4} \cdot 2^c$. Then, the probability of agreement is at least $\frac{1}{2} |C| \cdot B$.

**Proof.** If Alice outputs $a \in \{0, 1\}^k$ then either there exists a unique $v \in C$ such that $D(v) = a$ and $\langle v, X \rangle > t$, which happens with probability at most $\Pr[\langle v, X \rangle > t] \leq U$, or $\kappa_A(X) = a$, which happens with probability $2^{-k}$. The min-entropy guarantee follows.

For the agreement, fix $v \in C$. Define the event $E_v := A_v \land B_v \land C_v$ where $A_v := \{\langle v, X \rangle > t \land \langle v, Y \rangle > s\}$, $B_v := \{\exists w \neq v : \langle w, X \rangle > t\}$, and $C_v := \{\exists w \neq v : \chi(v) = \chi(w) \land \langle w, Y \rangle > s\}$. 


Note that the event $E_v$ ensures that both players output $D(v)$. By the union bound:

$$
\Pr[E_v] \geq \Pr[A_v \cdot (1 - \Pr[B_v \lor C_v | A_v])]
\geq \Pr[A_v \cdot \left(1 - \sum_{w \neq v} \Pr[(w, X) > t | A_v] - \sum_{w \neq v} 1\{\chi(w) = \chi(v)\} \cdot \Pr[(w, Y) > s | A_v]\right)
\geq B(1 - |C| \cdot \mathcal{T}_A - |C| \cdot 2^{-c} \cdot \mathcal{T}_B)
\geq \frac{1}{2} B,
$$

where the last two inequalities follow from the definition of $B$ and $\mathcal{T}$ and then invoking the premise of the lemma. Thus, the agreement probability is at least $\sum_v \Pr[E_v] \geq \frac{1}{2} |C| \cdot B$. □

As an illustration, we present an explicit one-way scheme for the BGS($\rho$) using an exponential number of samples. Let $k$ be a large enough constant and let $n = 2^k$. Let $\mathcal{C}$ consist of the $n$ standard basis vectors $\{e_i : i \in [n]\}$ in $\mathbb{R}^n$. Choose $t > 0$ so that the Gaussian tail probability $Q(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}$. Let $\rho \leq \phi \leq 1$ be arbitrary and set $s = \phi t$. (Choose $\phi = 1$ for zero communication.)

For the analysis, note that for each $i$, we have $\langle e_i, X \rangle = X_i$ and $\langle e_i, Y \rangle = Y_i$. Therefore, $\Pr[X_i > k] = Q(t)$ and so by Theorem 2.2.1, the min-entropy of Alice’s output is at least $- \log(Q(t) + 2^{-k}) \geq k - 1$.

We now analyze the agreement probability. To bound the bivariate tail, first by Proposition 2.1.4 (a), we have $t = \Theta(\sqrt{k})$. Let $\delta$ satisfy $\phi = \rho + \delta \sqrt{1 - \rho^2}$. Observe that $0 \leq \delta \leq 1$.

Applying Proposition 2.1.4 (b,c), we obtain:

$$
B = \min_{i \in [n]} \Pr[X_i > t, Y_i > \phi t] = L(t, \phi; \rho) \geq \frac{Q(t)^{1+\delta^2}}{\delta t + \Theta(1)} \geq \frac{Q(t)^{1+\delta^2}}{\delta \sqrt{k} + \Theta(1)} \quad \text{(2.1)}
$$

For $i \neq j$, the trivariate tail probability $\Pr[X_j > t \mid X_i > t, Y_i > \phi t] = \Pr[X_j > t] = Q(t)$, by independence of components of $(X, Y)$. Similarly, $\Pr[Y_j > \phi t \mid X_i > t, Y_i > \phi t] = Q(\phi t)$. Therefore:

$$
\mathcal{T}_A \leq Q(t) \quad \text{and} \quad \mathcal{T}_B \leq Q(\phi t). \quad \text{(2.2)}
$$

Now $Q(t) = \frac{1}{\sqrt{2\pi}} \cdot 2^{-k}$, so $|C| \cdot \mathcal{T}_A \leq \frac{1}{4}$. Next, $Q(\phi t) \leq Q(t)^{\phi^2}$, using Proposition 2.1.4 (d). Therefore, $\mathcal{T}_B \leq Q(t)^{\phi^2}$. If we choose $c \geq (1 - \phi^2)(k + 2)$, then it can be verified that $|C| \cdot \mathcal{T}_B \leq \frac{1}{4} \cdot 2^c$. This ensures that the conditions of Theorem 2.2.1 for agreement are satisfied.

By Theorem 2.2.1, the agreement probability is $\frac{1}{2} |C| \cdot B \geq 2^{-\delta^2 k / (\delta \sqrt{k} + \Theta(1))}$ and the scheme uses $O((1 - \phi^2)k)$ bits of communication. In particular, set $\phi = \rho$ and $\delta = 0$; we obtain an explicit one-way scheme with constant probability and $O((1 - \rho^2)k)$ bits of communication.
2.3 Efficient Scheme for BGS(\(\rho\))

In this section, we give a resource-efficient one-way scheme for BGS(\(\rho\)) with the optimal communication of \((1-\rho^2)k\) bits. More generally, the tradeoff between the communication and agreement probability is similar to the one achieved by the scheme presented in Section 2.2.

The analysis of the template given previously suggests the following scheme to reduce the sample complexity to \(n = \text{poly}(k)\): use a codebook such that the projections are only 3-wise independent. Unfortunately, this does not work since a multivariate Gaussian distribution is completely characterized by its first and second moments, so even pairwise independence would imply full independence! Instead, we use a codebook consisting of the following explicitly defined large family of nearly-orthogonal vectors in \(\mathbb{R}^n\) due to Tao [Tao13], who showed their near-orthogonality property using the Weil bound for curves.

Let \(p\) be a prime number and \(n = 2 \cdot p\). We identify \(\mathbb{R}^n\) with the complex vector space \(\mathcal{V}\) of functions from \(\mathbb{F}_p\) to \(\mathbb{C}\), where \(\mathbb{C}\) denotes the complex plane. Thus, \(v \in \mathcal{V}\) will also denote an element of \(\mathbb{R}^n\). With this identification, we have \(\langle v, w \rangle = \text{Re}(\sum_{x \in \mathbb{F}_p} v(x)w(x))\) for \(v, w \in \mathcal{V}\).

Let \(d\) be a positive integer. Let \(\omega := e^{2\pi i/p}\) denote the \(p\)-th root of unity. For every \(a \in \mathbb{F}_p^d\), let \(v_a \in \mathcal{V}\) be defined as \(v_a(x) = \frac{1}{\sqrt{p}} \cdot \omega^{a_1x_1 + \ldots + a_dx_d}\). We set \(\mathcal{C} := \{v_a : a \in \mathbb{F}_p^d\}\). Note that all the elements of \(\mathcal{C}\) have unit norm. The Weil bound for curves then implies that for every \(a \neq b \in \mathbb{F}_p^d\), we have that \(\left|\langle v_a, v_b \rangle\right| \leq (d-1)/\sqrt{p}\) [Wei48] (for a recent exposition see [KL11]).

Choose \(d = o(n^{1/4}/\sqrt{\log n})\) and \(k = d \cdot \log(n/2)\) in Tao’s construction. We use the same parameters \(t, s, \varphi\) and \(\delta\) for Protocol 1 as in the previous scheme described in Section 2.2.

By elliptical symmetry (Proposition 2.1.5), \((\langle v, X\rangle, \langle v, Y\rangle) \sim \text{BGS}(\rho)\), for every \(v \in \mathcal{C}\). Therefore, the bounds in Section 2.2 for the univariate and bivariate tails (see Equation (2.1)) also hold here. The key difference is in the analysis of the trivariate probabilities because we no longer have independence amongst the various pairs \((\langle v, X\rangle, \langle v, Y\rangle)\). This requires a new analysis of the conditional tails involving trivariate Gaussians whose covariances have a special structure. Below, we show that a slightly weaker bound than Equation (2.2): \(\mathcal{T}_A \leq Q(t) \cdot (1+o_n(1))\) and \(\mathcal{T}_B \leq Q(\varphi t) \cdot (1+o_n(1))\). Nevertheless, we can apply the same argument following Equation (2.2) in Section 2.2 (along with the min-entropy lower bound in Section 2.10.2) to get that: (a) the min-entropy at least \(k-1\); (b) the agreement probability is \(\gtrsim 2^{-\delta^2k}/(\delta \sqrt{k} + \Theta(1))\); and (c) the communication is \(O((1-\varphi^2)k)\) bits. In particular, with \(\varphi = \rho\) and \(\delta = 0\); we obtain the main result of this section, namely a resource-efficient one-way scheme using \(O((1-\rho^2) \cdot k)\) bits of communication.

It remains to prove that \(\mathcal{T}_A \leq Q(t) \cdot (1+o_n(1))\) and \(\mathcal{T}_B \leq Q(\varphi t) \cdot (1+o_n(1))\). Fix \(v \neq w \in \mathcal{C}\). The construction ensures that \(\left|\langle v, w \rangle\right| \leq \theta\) with \(\theta = (d-1)/\sqrt{p} = O(k/\sqrt{n} \cdot \log n)\). For \(k = o(n^{1/4} \cdot \sqrt{\log n})\), we have \(\theta = o_n(1)\).

Now observe that \((\langle w, X\rangle, \langle v, X\rangle, \langle v, Y\rangle)\) can be written as a linear transform on \((X, Y)\), so jointly they have the trivariate Gaussian distribution. Thus, their joint distribution is fully given by the first two moments. By stability, the marginals are standard normal and by elliptical symmetry, the covariances can be calculated as (i) \(\mathbb{E}[(\langle w, X\rangle)(\langle v, X\rangle)] = \langle w, v \rangle \leq \theta\),
(ii) \( \mathbb{E}[(\langle v, X \rangle)(\langle v, Y \rangle)] = \rho \), and (iii) \( \mathbb{E}[(\langle w, X \rangle)(\langle v, Y \rangle)] = \rho(\langle w, v \rangle) \leq \rho \). Observe that \( \langle w, Y \rangle, \langle v, Y \rangle, \langle v, X \rangle \) is also trivariate with an identical mean and covariance matrix.

**Lemma 2.3.1.** Let \((U, V, W)\) be a trivariate Gaussian with standard normal marginals and covariances \( \mathbb{E}[UV] = \sigma \), \( \mathbb{E}[VW] = \rho \), and \( \mathbb{E}[UW] = \sigma \rho \). Let \( r, r' \geq 0 \). Then, for all \( b \geq 1 \):

\[
\Pr[U > r | V > r, W > r'] \leq Q \left( \frac{1 - b \sigma}{\sqrt{1 - \sigma^2}} r \right) + \frac{Q(br)}{\Pr[V > r, W > r']}. 
\]

**Proof.** We have:

\[
\Pr[U > r | V > r, W > r'] = \frac{\Pr[U > r, V > r, W > r']}{\Pr[V > r, W > r']}. 
\]

For the numerator, we split the range of \( V \) into two intervals:

\[
\Pr[U > r, V > r, W > r'] = \Pr[U > r, V < r, W > r'] + \Pr[U > r, V > br, W > r']
\]

The second term is at most \( \Pr[V > br] = Q(br) \). For the first term, note that the covariance structure implies that \( U \) and \( W \) are independent conditioned on \( V \), so we can write \( U = \sigma V + \sqrt{1 - \sigma^2}Z \), where \( Z \sim \mathcal{N}(0,1) \) is independent of \((V, W)\). The event \( \{U > r\} \) can be rewritten as \( \{Z > \frac{r - \sigma V}{\sqrt{1 - \sigma^2}}\} \) which under the assumption \( \{V \leq br\} \) implies that \( \{Z > ar\} \) where \( a := \frac{1 - b \sigma}{\sqrt{1 - \sigma^2}} \). By independence:

\[
\Pr[U > r, V < r, W > r'] \leq \Pr[Z > ar]\Pr[r < V \leq br, W > r'] \leq Q(ar)\Pr[V > r, W > r']
\]

Applying these bounds in Equation (2.3) finishes the proof. \( \square \)

Apply Lemma 2.3.1 to the triples \((\langle w, X \rangle, \langle v, X \rangle, \langle v, Y \rangle)\) with \( r := t \), \( r' := \varphi t \) and \((\langle w, Y \rangle, \langle v, Y \rangle, \langle v, X \rangle)\) with \( r := \varphi t \), \( r' := t \). In both cases, \( \sigma := \langle v, w \rangle \leq \theta \). Since \( Q(\cdot) \) is decreasing:

\[
\Pr[\langle w, X \rangle > t | \langle v, X \rangle > t, \langle v, Y \rangle > \varphi t] \leq Q \left( \frac{1 - b \theta}{\sqrt{1 - \theta^2}} t \right) + \frac{Q(bt)}{L(t, \varphi; \rho)}, \quad \forall b \geq 1. 
\]

\[
\Pr[\langle w, Y \rangle > \varphi t | \langle v, Y \rangle > \varphi t, \langle v, X \rangle > t] \leq Q \left( \frac{1 - b \theta}{\sqrt{1 - \theta^2}} \varphi t \right) + \frac{Q(b\varphi t)}{L(t, \varphi; \rho)}, \quad \forall b \geq 1. 
\]

Set \( b := 2/\phi \). By Proposition 2.1.4 (a), \( Q(b\varphi t) \lesssim e^{-b^2\varphi^2 t^2/2} = e^{-2\varphi t} \) and \( Q(t) \gtrsim e^{-t^2/2}(t + \lambda_0) \). Because \( t = \Theta(\sqrt{k}) \), for large enough \( k \), we have \( Q(b\varphi t) \lesssim Q(t)^3 e^{-t^2/2}(t + \lambda_0)^3 = Q(t)^3 o_n(1) \).

Using this bound and Proposition 2.1.4 (c,d), we obtain:

\[
Q(bt) \leq Q(b\varphi t) \leq Q(t)^3 \cdot o_n(1) \leq L(t, \varphi; \rho)Q(t) \cdot o_n(1) \leq L(t, \varphi; \rho)Q(\varphi t) \cdot o_n(1)
\]

Thus, the second term in Equation (2.4) (respectively Equation (2.5)) is at most \( Q(t) \cdot o_n(1) \).
(respectively $Q(\varphi t) \cdot o_n(1)$).

For the first terms on the right side of Equation (2.4) and Equation (2.5), let $a := \frac{1-b\theta}{\sqrt{1-a^2}}$. Note that $a \leq 1$. Now $Q(at) \leq Q(t) a^2$ by Proposition 2.1.4(d). We calculate
\[
1 - a^2 = \left(\frac{2b/(1+b^2)\theta}{1-a^2}\right) \leq 4b\theta, \text{ since } \theta \ll 2b/(1+b^2).
\]
For the choice of $d$, we have $kb\theta = o_n(1)$. Thus, $Q(at)/Q(t) \leq Q(t) a^2 \leq 2^{k(1-a^2)} \leq 2^{4kb\theta} = 2o_n(1) = 1 + o_n(1)$.

By Lemma 2.6.3, $Q(at)/Q(t)$ is increasing in $t$, so $Q(at)/Q(t) \leq Q(at)/Q(t) \leq Q(at)/Q(t) + o_n(1)$. Thus, the first term in Equation (2.4) (respectively Equation (2.5)) is at most $Q(t) \cdot (1 + o_n(1))$ (respectively $Q(\varphi t) \cdot (1 + o_n(1))$). Combine the above bounds for the two terms in Equations (2.4) and (2.5) to complete the analysis. This completes the proofs of Theorem 2.1.1 and Theorem 2.1.2 for the BGS source.

Remark 2.3.2. We modify the above resource-efficient scheme that uses $c = (1 - \rho^2) \cdot k$ bits of communication to generate secret keys. Assume without loss of generality that codewords within the same color class are encoded with the same prefix of $c$ bits. Now Alice just outputs the $(k-c = \rho^2 \cdot k)$-bit suffix of her output. We briefly sketch the analysis as follows. Using the min-entropy property as well as a similar lower bound on the probability that Alice outputs a particular key (which essentially follows from the same bounds on bivariate tails used above), it can be shown that the communicated bits are nearly uniform as well and that the suffix of the output is nearly uncorrelated with the prefix. This ensures the secrecy of the key from the eavesdropper.

\subsection*{2.4 Efficient Scheme for DSBS(\(\rho\))}

We give a resource-efficient one-way scheme for DSBS(\(\rho\)) with the optimal communication of $(1 - \varphi^2) \cdot k$ using the template of Protocol 1. It is based on dual-BCH codes which can be seen as finite field analogues of the nearly-orthogonal vectors used in Section 2.3. It is more natural here, but still equivalent, to work with $\{0,1\}^n$ instead of $\{\pm 1\}^n$ and the Hamming distance $\Delta$ instead of the inner product over $\mathbb{R}$. We start by recalling the definition and basic properties of dual-BCH codes.

Dual-BCH Codes. Let $d,m \geq 1$ be integers satisfying $2 \cdot d - 2 < 2^{m/2}$. Consider the Reed-Solomon code $C_{RS}$ which is obtained by evaluating all univariate polynomials of degree at most $2^m - 2 \cdot d - 1$ over all non-zero elements of the finite field $\mathbb{F}_{2^m}$. The BCH code is then defined as $C_{BCH(d,m)} \triangleq C_{RS} \cap \mathbb{F}_{2^m}^{2^m-1}$. Specifically, the BCH code is the subset of all binary codewords of the Reed-Solomon code. Then, the dual-BCH code is defined to be the dual code $C_{dBCH(d,m)} \triangleq (C_{BCH(d,m)})^\perp$. Namely, a vector $v \in \mathbb{F}_{2^m}^{2^m-1}$ is a dual-BCH codeword if and only if $\langle v, v' \rangle = 0$ for all $v' \in C_{BCH(d,m)}$ (where here the inner product is over $\mathbb{F}_2$). The message length of $C_{dBCH(d,m)}$ is equal to $d \cdot m$, and hence $|C_{dBCH(d,m)}| = 2^{d \cdot m}$. Dual-BCH codes are known to be “unbiased” codes, a fact that again follows from the Weil bound for curves [Wei48].
Theorem 2.4.1 (Weil-Carlitz-Uchiyama Bound [MS77]). Let \(d, m \geq 1\) be integers satisfying 
\(2 \cdot d - 2 < 2^{m/2}\). Then, for every non-zero codeword \(v \in C_{d BCH}(d, m)\), we have that
\[2^{m-1} - (d - 1) \cdot 2^{m/2} \leq \wt(v) \leq 2^{m-1} + (d - 1) \cdot 2^{m/2}.
\]

We are now ready to give our efficient scheme for DSBS(\(\rho\)). Let \(C_{d BCH} = C_{d BCH}(d, m)\) be the dual-BCH code with parameters \(m = \log(n+1)\) and \(d\) being any polynomial in \(n\) that satisfies \(d = o(n^{1/4}/\sqrt{\log n})\). Then, \(|C_{d BCH}| = 2^k\) where \(k = d \cdot \log(n+1)\) is a polynomial in \(n\). Let \(\mathcal{C}\) be an arbitrary subset of \(C_{d BCH}\) of size \(2^k = (\gamma \cdot n)\), where \(\gamma > 0\) is a sufficiently large absolute constant to be chosen later on. We denote \(\mathcal{C} = \{v_a : a \in \{0, 1\}^{k'}\}\). We set \(r = n/2 - t\sqrt{n}/2\) where \(t > 0\) satisfies \(Q(t) = (1/4) \cdot 2^{-k}\). Similar to before, let \(\rho \leq \varphi \leq 1\) so that the communication is \(O((1 - \varphi^2)k)\) bits. Recall that \(\delta\) satisfies \(\varphi = \rho + \delta \sqrt{1 - \rho^2}\).

In the following, we prove the appropriate uni-, bi- and trivariate tail bounds for the scheme. These are stated in Proposition 2.4.2 and lemmas 2.4.3 and 2.4.4. The proof follows the same structure that was used for BGS(\(\rho\)). It requires some bounds on (correlated) binomial sums proved in Section 2.7. By incorporating them into Theorem 2.2.1, we obtain the desired performance of the scheme. Let \((X, Y) \sim DSBS(\rho)^{\otimes n}\).

Proposition 2.4.2. For any \(u \in \mathbb{R}\) (possibly depending on \(n\)), \(\Pr[|\wt(X) - n/2| \geq u\sqrt{n}/2] \leq \text{poly}(n) \cdot Q(u)\).

Lemma 2.4.3. For every \(a \in \{0, 1\}^{k'}\), \(\Pr[\Delta(v_a, X) \leq r, \Delta(v_a, Y) \leq r'] \geq \frac{1}{\Theta(n^2)} \cdot 2^{-k} \cdot 2^{-k\delta^2}.
\)

Proof. Follows from Proposition 2.7.3 and Proposition 2.7.5.

Lemma 2.4.4. Let \(v, v' \in \{0, 1\}^n\) satisfy \(|\Delta(v, v') - n/2| \leq \theta \cdot n/2\), where \(\theta = O(k/(\sqrt{n} \cdot \log n))\). Then:
\[
\begin{align*}
\Pr[\Delta(v', X) \leq r | \Delta(v, X) \leq r, \Delta(v, Y) \leq r'] &\leq O(n) \cdot Q(t) \\
\Pr[\Delta(v', Y) \leq r' | \Delta(v, X) \leq r, \Delta(v, Y) \leq r'] &\leq O(n) \cdot Q((1 - \varphi)t)
\end{align*}
\]

Proof. Let \(\ell = n/2 - \theta \cdot n/2\). Without loss of generality, we assume that \(v = 0^n\) is the
all-zeros vector and that \( v' = 1^t 0^{n-t} \). Then,

\[
\Pr[\Delta(v', X) \leq r \mid \Delta(v, X) \leq r, \Delta(v, Y) \leq r'] \\
= \Pr[\Delta(v', X) \leq r \mid \text{wt}(X) \leq r, \text{wt}(Y) \leq r'] \\
= \frac{\Pr[\Delta(v', X) \leq r, \text{wt}(X) \leq r, \text{wt}(Y) \leq r']}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']}
\]

\[
= \frac{1}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']}
\cdot \sum_{r_1=0}^r \sum_{r_2=0}^{r'} \sum_{r_3=0}^{r'} \Pr[\Delta(v', X) = r_1, \text{wt}(X) = r_2, \text{wt}(Y) = r_3]
\]

\[
= \frac{1}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']}
\cdot \sum_{r_2=0}^r \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \cdot \Pr[\Delta(v', X) = r_1, \text{wt}(X) = r_2] \\
= \frac{1}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']}
\cdot \sum_{r_2=0}^r \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \cdot \Pr[\Delta(v', X) \leq r, \text{wt}(X) = r_2], \tag{2.6}
\]

where the penultimate equality follows from the fact that \( \Delta(v', X) - \text{wt}(X) - \text{wt}(Y) \) is a Markov chain.

For every non-negative integer \( t_2 \) satisfying \( t_2 = o(n^{1/4}) \) and \( \theta \cdot t \cdot t_2 = o_n(1) \), we have that

\[
\Pr[\Delta(v', X) \leq r \mid \text{wt}(X) = n/2 - t_2 \sqrt{n}/2] \\
= \sum_{a=0}^{a_{\text{max}}} \psi(a) \tag{A}
\leq (a_{\text{max}} + 1) \cdot \psi(a_{\text{max}}) \tag{B}
\leq O(n) \cdot \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot e^{-\frac{r^2}{2}} \tag{C}
\leq O(n) \cdot Q(t), \tag{2.7}
\]

where (A) follows from Proposition 2.7.6, (B) from Proposition 2.7.7 and the fact that \( \theta = o_n(1) \), and (C) from Proposition 2.1.4 (a) and the facts that \( t = \Theta(\sqrt{k}) \) and \( k \leq n \). Note that by assumption \( \theta = O(k/(\sqrt{n} \cdot \log n)) \). Thus, for any \( k = o(n^{1/4} \cdot \sqrt{\log n}) \), there exists a
function \( \nu(t, \theta) = \omega_n(1) \) satisfying \( \nu(t, \theta) = o_n(\min(n^{1/4}, 1/(t \cdot \theta))) \) and

\[
\Theta(n^2) \cdot 2^{k + k \delta^2} \cdot \exp(-\nu(t, \theta)^2) \leq Q(t).
\]  (2.8)

We fix such a function \( \nu(t, \theta) \) and set \( \tau(t, \theta) \overset{\Delta}{=} n/2 - \nu(t, \theta) \sqrt{n}/2 \). Equation (2.6) now becomes:

\[
\Pr[\Delta(v', X) \leq r \mid \Delta(v, X) \leq r, \Delta(v, Y) \leq r'] = \frac{1}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']} \cdot (\alpha + \beta),
\]  (2.9)

where

\[
\alpha \overset{\Delta}{=} \sum_{r_2=\tau(t, \theta)}^{r} \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \cdot \Pr[\Delta(v', X) \leq r \mid \text{wt}(X) = r_2],
\]

and

\[
\beta \overset{\Delta}{=} \sum_{r_2=0}^{\tau(t, \theta)} \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \cdot \Pr[\Delta(v', X) \leq r \mid \text{wt}(X) = r_2].
\]

Using Equation (2.7) and the fact that \( \nu(t, \theta) = o(\min(n^{1/4}, 1/(t \cdot \theta))) \), we get that

\[
\alpha \leq \sum_{r_2=\tau(t, \theta)}^{r} \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \cdot O(n) \cdot Q(t) \leq O(n) \cdot Q(t) \cdot \Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r'],
\]  (2.10)

We also have that

\[
\beta \leq \sum_{r_2=0}^{\tau(t, \theta)} \sum_{r_3=0}^{r'} \Pr[\text{wt}(X) = r_2, \text{wt}(Y) = r_3] \leq \Pr[\text{wt}(X) \leq \tau(t, \theta)] \leq \exp(-\nu(t, \theta)^2),
\]  (2.11)

where the last inequality uses the fact that \( \nu(t, \theta) = \omega_n(1) \) and follows from Proposition 2.4.2 and Proposition 2.1.4 (a). Combining Equation (2.9), Equation (2.10) and Equation (2.11), we get that

\[
\Pr[\Delta(v', X) \leq r \mid \Delta(v, X) \leq r, \Delta(v, Y) \leq r'] \leq O(n) \cdot Q(t) + \frac{\exp(-\nu(t, \theta)^2)}{\Pr[\text{wt}(X) \leq r, \text{wt}(Y) \leq r']} \leq O(n) \cdot Q(t) + \Theta(n^2) \cdot 2^{k} \cdot 2^{-k \delta^2} \cdot \exp(-\nu(t, \theta)^2) \leq O(n) \cdot Q(t),
\]

where the second inequality follows from Proposition 2.7.3 and Proposition 2.7.5, and the third inequality follows from the fact that \( \nu(t, \theta) \) satisfies Equation (2.8). This completes the proof of the first part of the lemma. The proof of the second part follows along the same lines. \( \square \)
We note that the above bounds imply the desired result for agreement probabilities up to $1/\text{poly}(k)$. The result also holds for constant agreement probability. The main idea is to combine the constant agreement scheme for the Gaussian source along with a multi-dimensional Berry-Esseen Theorem (e.g., Theorem 67 of [MORS10]).\footnote{Since we are dealing with constant error probabilities, the additive error from the Berry-Esseen theorem is negligible.}

The min-entropy guarantee for the above schemes follows from Section 2.10.1.

2.5 Information Complexity and Common Randomness

In this section, we show an intimate relationship between the achievable regions for amortized common randomness generation and the internal and external information costs of communication protocols, two well-studied notions in theoretical computer science. For any random variable $X$, let $H(X)$ denote its Shannon entropy. We say that a triple $(H, R_1, R_2)$ of non-negative real numbers is $r$-achievable for a distribution $\mu$ if for every $\epsilon > 0$ there exists an $r$-round common randomness scheme $\Pi$ with inputs $(X^n, Y^n) \sim \mu^{\otimes n}$ for some $n = n(\epsilon)$ where $n \to \infty$ as $\epsilon \to 0$, such that the following holds: let $M_t$ denote the message sent in round $t$ in $\Pi$, and let $K_A$ (respectively, $K_B$) denote the output of Alice (respectively, Bob). Then, (1) $\sum_{t \text{ odd}} H(M_t) \leq (R_1 + \epsilon)n$, (2) $\sum_{t \text{ even}} H(M_t) \leq (R_2 + \epsilon)n$, (3) $H(K_A), H(K_B) \geq (H - \epsilon)n$, (4) $K_A$ and $K_B$ both belong to a domain of size $cn$ for some absolute constant $c$ independent of $\epsilon$ and $n$, and (5) $\Pr[K_A \neq K_B] \leq \epsilon$. We note that the min-entropy guarantee in our basic definition is stronger than the combination of parts (3) and (4) here.

**Definition 2.5.1.** Let $P$ be a two-player randomized communication protocol with both public and private coins and let $R_{\text{pub}}$ denote the public randomness. With a slight abuse in notation, given $(X, Y) \sim \mu$, let $P$ also denote the transcript of the protocol on input $(X, Y)$. Define the following measures for the protocol with respect to $\mu$: (i) the external information cost $\text{IC}^\text{ext}(P)$ equals $I(X, Y; P \mid R_{\text{pub}})$; (ii) the marginal internal information cost $\text{IC}^\text{int}_A(P)$ for Alice equals $I(X; P \mid YR_{\text{pub}})$ and analogously $\text{IC}^\text{int}_B(P) = I(Y; P \mid XR_{\text{pub}})$ for Bob. The (total) internal information cost equals the sum of the two marginal costs.

We now characterize the achievable region for a fixed source distribution $\mu$ in terms of internal and external information costs of protocols with respect to $\mu$.

**Converse.** We extend the ideas present in several works, e.g. [Kas85, AC98, LCV16]. We need the following direct-sum property (lemma 2.5.3 below) for information costs of randomized protocols that we crucially use in our analysis. This property differs from the known direct-sum results in that it simultaneously bounds the internal and external information costs of the single-coordinate protocol. Its proof uses the following tool.
Proposition 2.5.2 ([AC98, Lemma 4.1]). Let $S, T, X^n, Y^n$ be arbitrary random variables. Then:

$$I(X^n; S \mid T) - I(Y^n; S \mid T) = \sum_{j=1}^{n} I(X_j; S \mid X_j^{j-1}Y_{j+1}^n T) - I(Y_j; S \mid X_j^{j-1}Y_{j+1}^n T).$$

Proof. We have by telescoping:

$$I(X^n; S \mid T) - I(Y^n; S \mid T) = \sum_{j=1}^{n} I(X^jY_{j+1}^n; S \mid T) - I(X_j^{j-1}Y_{j+1}^n; S \mid T). \quad (2.12)$$

By the chain rule for mutual information, for each $j \in [n]$, we have that

$$I(X^jY_{j+1}^n; S \mid T) = I(X_j^{j-1}Y_{j+1}^n; S \mid T) + I(X_j; S \mid X_j^{j-1}Y_{j+1}^n T)$$

and

$$I(X_j^{j-1}Y_{j+1}^n; S \mid T) = I(X_j^{j-1}Y_{j+1}^n; S \mid T) + I(Y_j; S \mid X_j^{j-1}Y_{j+1}^n T)$$

The proposition now follows by substituting the last two equations in Equation (2.12). \qed

Lemma 2.5.3 (Direct sum). Fix a distribution $\mu$ and an $r$-round randomized protocol $\Pi$ with inputs $(X^n, Y^n) \sim \mu^\otimes n$. Then, there exists an $r$-round randomized protocol $P$ with inputs $(X, Y) \sim \mu$ such that (a) $\text{IC}^\text{int}_A(\Pi) = n \cdot \text{IC}^\text{int}_A(P)$, (b) $\text{IC}^\text{int}_B(\Pi) = n \cdot \text{IC}^\text{int}_B(P)$, and (c) $\text{IC}^\text{ext}(\Pi) \leq n \cdot \text{IC}^\text{ext}(P)$.

Proof. For ease of presentation, we suppress the public randomness of $\Pi$ in the expressions appearing in the proof below. Let $M_t$ be the message sent in $\Pi$ during round $t \in [r]$; set $M_{r+1} := \emptyset$. We will be using the following properties of $\Pi$:

I. For every odd $t \leq r$, $I(Y^n; M_t \mid X^n M_t^{t-1}) = I(X^n; M_{t+1} \mid Y^n M_t) = 0$.

II. For all $j \in [n]$ and odd $t \leq r$, $I(Y_j; M_t \mid X^n Y_j Y_{j+1}^n M_t^{t-1}) = I(X_j; M_{t+1} \mid X_j^{j-1}Y_{j+1}^n M_t) = 0$.

This can also be shown, see, e.g., [Kas85, Eqns. 3.10–3.13].

We present the argument for the marginal internal information cost for Alice; a similar argument can be carried out for Bob’s case as well. Observe that:

$$\text{IC}^\text{int}_A(\Pi) = I(X^n; M^r \mid Y^n) = \sum_{t \leq r} I(X^n; M_t \mid Y^n M_t^{t-1}) = \sum_{t \text{ odd}} I(X^n; M_t \mid Y^n M_t^{t-1}), \quad (2.13)$$

by item (I) above. Fix an odd $t$ in the above sum. Again by item (I) above:

$$I(X^n; M_t \mid M_t^{t-1}) = I(X^n Y^n; M_t \mid M_t^{t-1}) = I(Y^n; M_t \mid M_t^{t-1}) + I(X^n; M_t \mid Y^n M_t^{t-1}), \quad (2.14)$$

38
and therefore,

\[ I(X^n; M_t \mid Y^n M^{t-1}) = I(X^n; M_t \mid M^{t-1}) - I(Y^n; M_t \mid M^{t-1}) \]

\[ \overset{(a)}{=} \sum_{j=1}^{n} I(X_j; M_t \mid X_j^{-1} Y^n_{j+1} M^{t-1}) - I(Y_j; M_t \mid X_j^{-1} Y^n_{j+1} M^{t-1}) \]

\[ = \sum_{j=1}^{n} I(X_j; M_t \mid Y_j X_j^{-1} Y^n_{j+1} M^{t-1}) \]

\[ = \sum_{j=1}^{n} I(X_j; M_{t+1} \mid Y_j X_j^{-1} Y^n_{j+1} M^{t-1}) \]

(2.15)

where (a) follows from proposition 2.5.2, and each of the last two equalities follows from the chain rule and by invoking item (II). We now substitute eq. (2.15) in eq. (2.13), and sum over all odd \( t \):

\[ \text{IC}_{A}^{\text{int}}(\Pi) = I(X^n; M^r \mid Y^n) = \sum_{t \text{ odd}} \sum_{j=1}^{n} I(X_j; M_{t+1} \mid Y_j X_j^{-1} Y^n_{j+1} M^{t-1}) \]

\[ = \sum_{j=1}^{n} I(X_j; M^r \mid Y_j X_j^{-1} Y^n_{j+1}) = n \cdot I(X_J; M^r \mid Y_J X_J^{-1} Y^n_{J+1}, J), \]

(2.16)

using the chain rule and then defining \( J \) to be uniform over \([n]\) and independent of all the other random variables. Similarly for Bob:

\[ \text{IC}_{B}^{\text{int}}(\Pi) = n \cdot I(Y_J; M^r \mid X_J X_J^{-1} Y^n_{J+1}, J). \]

(2.17)

We claim that the right side of eqs. (2.16) and (2.17) are respectively the marginal internal information costs for Alice and Bob in some protocol \( P \) with inputs \((X, Y) \sim \mu\). Specifically, on input pair \((X, Y)\), the protocol \( P \) simulates the protocol \( \Pi \) by setting \( X_J \coloneqq X \) and \( Y_J \coloneqq Y \), and associating the public randomness with \( J \), \( X_J^{-1} \), and \( Y^n_{J+1} \). Item (II) above ensures that the messages in protocol \( P \) can be generated by the players using private randomness.

It remains to bound the external information cost of \( P \). Observe that \( \text{IC}_{A}^{\text{ext}}(P) \) equals

\[ I(X_J, Y_J; M^r \mid X_J^{-1} Y^n_{J+1} J) = I(Y_J; M^r \mid X_J^{-1} Y^n_{J+1} J) + I(X_J; M^r \mid Y_J X_J^{-1} Y^n_{J+1} J). \]

(2.18)

The second term in eq. (2.18) above equals \( \frac{1}{n} \cdot I(X^n; M^r \mid Y^n) \) via eq. (2.16). For the first term, using the independence of coordinates,

\[ I(Y_J; M^r \mid X_J^{-1} Y^n_{J+1} J) = I(Y_J; M^r X^{-1} \mid Y^n_{J+1} J) \geq I(Y_J; M^r \mid Y^n_{J+1} J) = \frac{1}{n} \cdot I(Y^n; M^r), \]

where we expand over \( J \) and use the chain rule. Combining the bounds for the two terms,
we conclude:

\[ n \cdot \text{IC}^{\text{ext}}(P) \geq I(Y^n; M^r) + I(X^n; M^r \mid Y^n) = I(X^nY^n; M^r) = \text{IC}^{\text{ext}}(\Pi). \]

**Theorem 2.5.4.** If a tuple \((H, R_1, R_2)\) is \(r\)-achievable then for every \(\epsilon > 0\) there exists a randomized \(r\)-round protocol whose marginal internal information cost for Alice (respectively Bob) with respect to the distribution \(\mu\) is at most \(R_1 + O(\epsilon)+1/n\) (respectively \(R_2+O(\epsilon)+1/n\)) and whose external information cost is at least \(H - \epsilon\).

**Proof.** Fix \(\epsilon > 0\). Let \(n\) be such that there is an \(r\)-round protocol for common randomness generation \(\Pi\) on inputs \((X^n, Y^n) \sim \mu^{\otimes n}\). Let \(M_t\) denote the message sent in round \(t\) in \(\Pi\). Let \(K_A\) (respectively, \(K_B\)) denote the output of Alice (respectively, Bob). We have (1) \(\sum_{t \text{ odd}} H(M_t) \leq (R_1 + \epsilon)n\), (2) \(\sum_{t \text{ even}} H(M_t) \leq (R_2 + \epsilon)n\), (3) \(H(K_A), H(K_B) \leq (H - \epsilon)n\), (4) \(K_A\) and \(K_B\) both belong to a domain of size \(cn\) for some absolute constant \(c\) (independent of \(\epsilon\) and \(n\)), and (5) \(\Pr[K_A \neq K_B] \leq \epsilon\).

Consider the case where \(r\) is odd (the other case can be handled similarly) and define a new protocol \(\Pi'\) where Alice also sends \(K_A\) to Bob along with the last message. The number of rounds is still \(r\). Applying Lemma 2.5.3, there exists an \(r\)-round randomized protocol \(P\) with inputs \((X, Y) \sim \mu\) such that \(\text{IC}^{\text{int}}(\Pi') = n \cdot \text{IC}^{\text{int}}(P)\) and \(\text{IC}^{\text{int}}(\Pi) = n \cdot \text{IC}^{\text{int}}(P)\). Now since \(\Pi'\) depends only on \(X^n\) and \(Y^n\), we have that \(\text{IC}^{\text{int}}(\Pi') = I(X^n; M^rK_A \mid Y^n) = I(X^n; M^r \mid Y^n) + I(X^n; K_A \mid Y^nM^r)\). Because \(X^n \perp M_t \mid Y^nM^{t-1}\) for each even round \(t\), by the chain rule, the first term equals

\[ \sum_t I(X^n; M_t \mid Y^nM^{t-1}) = \sum_{t \text{ odd}} I(X^n; M_t \mid Y^nM^{t-1}) \leq \sum_{t \text{ odd}} H(M_t) \leq (R_1 + \epsilon)n. \]

The second term is at most \(H(K_A \mid Y^nM^r)\). Now \(K_B\) is determined by \(Y^n\) and \(M^r\) and \(\Pr[K_A \neq K_B] \leq \epsilon\), so by Fano’s inequality, \(H(K_A \mid Y^nM^r) \leq \epsilon cn + 1\). Therefore, \(\text{IC}^{\text{int}}(P) \leq R_1 + \epsilon(1 + c) + 1/n\). For Bob, the analysis is similar and even simpler because his messages are unchanged (from \(\Pi\) to \(\Pi'\)) so \(\text{IC}^{\text{int}}(P) \leq R_2 + \epsilon\). (The bound stated in the lemma is weaker because Fano’s inequality is used when \(r\) is even.) Finally, apply Lemma 2.5.3 to bound the external information cost of \(P\) as

\[ n \cdot \text{IC}^{\text{ext}}(P) \geq \text{IC}^{\text{ext}}(\Pi') = I(X^nY^n; M^rK_A) = H(M^rK_A) - H(M^rK_A \mid X^nY^n) = H(M^rK_A). \]

But \(H(M^rK_A) \geq H(K_A) \geq (H - \epsilon)n\), so the desired bound follows.

**Achievability.** In [LCV16], a sufficient condition using Markov chains on auxiliary random variables is given to the existence of an interactive common randomness scheme. To fulfill this condition, their construction uses a random encoding argument. We connect these conditions to the existence of an \(r\)-round communication protocol with the appropriate information costs.

**Proposition 2.5.5 ([LCV16]).** Let \((X, Y) \sim \mu\). Suppose there exist auxiliary random variables \(U_1, U_2, \ldots, U_r\) for some \(r\) in some joint probability space with \(X\) and \(Y\) where the
marginal distribution of \((X, Y)\) is \(\mu\) satisfying the following:

1. For every odd \(t\), \(Y \perp U_t \mid Xu^{t-1}\) and for every even \(t\), \(X \perp U_{t+1} \mid Yu^t\).
2. \(\sum_{t \text{ odd}} I(X; U_t \mid U^{t-1}) + \sum_{t \text{ even}} I(Y; U_t \mid U^{t-1}) \geq H\).
3. \(\sum_{t \text{ odd}} I(X; U_t \mid U^{t-1}) - \sum_{t \text{ odd}} I(Y; U_t \mid U^{t-1}) \leq R_1\).
4. \(\sum_{t \text{ even}} I(Y; U_t \mid U^{t-1}) - \sum_{t \text{ even}} I(X; U_t \mid U^{t-1}) \leq R_2\).

Then, there exists an \(r\)-round interactive common randomness generation scheme \(\Pi(X^n, Y^n)\) using \(n\) i.i.d. samples as input where Alice sends at most \(R_1n\) bits, Bob sends at most \(R_2n\) bits and the entropies of their outputs are each at least \(Hn\) bits where the agreement probability tends to 1 as \(n \to \infty\).

**Theorem 2.5.6.** If there exists an \(r\)-round randomized protocol with inputs \((X, Y) \sim \mu\) whose marginal internal information cost for Alice (respectively, Bob) is at most \(R_1\) (respectively, \(R_2\)) and whose external information cost is at least \(H\), then \((H, R_1, R_2)\) is \(r\)-achievable.

**Proof.** Let \(P\) be a randomized protocol with inputs \((X, Y) \sim \mu\) whose marginal internal information cost for Alice (respectively, Bob) is at most \(R_1\) (respectively, \(R_2\)) and whose external information cost is at least \(H\). Without loss of generality, we assume that \(P\) uses no public randomness. For every \(t \in [r]\), let \(U_t\) denote the message sent in \(P\) during round \(r\). We claim that the \(U_t\)'s satisfy the conditions in Proposition 2.5.5. First, note that the conditional independencies given in item 1 of Proposition 2.5.5 are equivalent to the message structure of an \(r\)-round randomized protocol, and are thus satisfied by the \(U_t\)'s.

For every odd \(t\), by item 1, \(I(Y; U_t \mid Xu^{t-1}) = 0\), so

\[I(X; U_t \mid U^{t-1}) = I(XY; U_t \mid U^{t-1}) = I(Y; U_t \mid U^{t-1}) + I(X; U_t \mid YU^{t-1}).\]

Therefore, \(I(X; U_t \mid YU^{t-1}) = I(X; U_t \mid U^{t-1}) - I(Y; U_t \mid U^{t-1})\). By the chain rule,

\[I_{\text{int}}^A(P) = I(X; U^r \mid Y) = \sum_t I(X; U_t \mid YU^{t-1}) = \sum_{t \text{ odd}} I(X; U_t \mid U^{t-1}) - I(Y; U_t \mid U^{t-1}) \leq R_1,
\]

via item 1 where we used \(I(X; U_t \mid YU^{t-1}) = 0\) for every even \(t\). Using the given assumption that \(I_{\text{int}}^A(P) \leq R_1\), we deduce that the \(U_t\)'s satisfy item 3 of Proposition 2.5.5. A similar argument using the given assumption that \(I_{\text{int}}^B(P) \leq R_2\) implies that the \(U_t\)'s satisfy item 4 of Proposition 2.5.5.

Applying a similar reasoning, we also obtain that:

\[\sum_{t \text{ odd}} I(X; U_t \mid U^{t-1}) + \sum_{t \text{ even}} I(Y; U_t \mid U^{t-1}) = \sum_{t \text{ odd}} I(XY; U_t \mid U^{t-1}) + \sum_{t \text{ even}} I(XY; U_t \mid U^{t-1})
\]

\[= I(XY; U^r)
\]

\[= I_{\text{ext}}(P).
\]
The given assumption that $IC^{\text{ext}}(P) \geq H$ now implies that the $U_i$’s satisfy item 2 of Proposition 2.5.5. Therefore, we conclude that $(H, R_1, R_2)$ is $r$-achievable.

Combining Theorem 2.5.4 and Theorem 2.5.6, we obtain the the formal version of Theorem 2.1.3.

**Theorem 2.5.7.** Let $\Gamma_r$ denote the supremum over all $r$-round randomized protocols $\Pi$ of the ratio of the external information cost to the internal information cost of $\Pi$ with respect to $\mu$. Then, $\Gamma_r$ equals the supremum of $H/(R_1 + R_2)$ such that $(H, R_1, R_2)$ is $r$-achievable for $\mu$.

### 2.6 Properties of Bivariate Gaussian Distribution

**Proposition 2.6.1** (Elliptical symmetry, Proposition 2.1.5 restated). Let $(X, Y) \sim \text{BGS}(\rho)^{\otimes n}$ and $v, w \in \mathbb{R}^n$ have unit norm. Then, $(\langle v, X \rangle, \langle w, Y \rangle) \sim \text{BGS}(\rho(\langle v, w \rangle))$.

*Proof.* Since $(\langle v, X \rangle, \langle w, Y \rangle)$ is a linear transform of $(X, Y)$, it has a bivariate Gaussian distribution. Thus, we only need to determine the first and second moments. Since $v$ and $w$ have unit-norm, by stability, the marginals are standard normal. Finally, we verify that their covariance is $\rho(\langle v, w \rangle)$:

$$
\mathbb{E}[\langle v, X \rangle \langle w, Y \rangle] = \sum_{i,j=1}^{n} v(i)w(j) \cdot \mathbb{E}[X_i \cdot Y_j] = \rho \sum_{i=1}^{n} v(i)w(i) = \rho(\langle v, w \rangle).
$$

### 2.6.1 Tail Bounds for Gaussians

The following bounds are well-known; using Duembgen’s approach [Due10], we prove them below for the sake of completeness. Let $\lambda(t) := \frac{\phi(t)}{Q(t)}$ denote the inverse Mills ratio, i.e., the ratio of the density function to the tail probability of a standard normal random variable. Let $\lambda_0 := \lambda(0) = \sqrt{\frac{2}{\pi}}$.

**Lemma 2.6.2.** For all $t \geq 0$, $\max\{t, \lambda_0^2 \cdot t + \lambda_0\} \leq \lambda(t) \leq t + \min\{1/t, \lambda_0\}$. Equality holds only at $t = 0$.

*Proof.* For all $t \geq 0$ and any function $\alpha: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\lim_{t \rightarrow \infty} \alpha(t) = \infty$, let

$$
f_\alpha(t) := \frac{\phi(t)}{\alpha(t)} - Q(t),
$$

so that $\lim_{t \rightarrow \infty} f_\alpha(t) = 0$. Observing that $Q'(t) = -\phi(t)$ and $\phi'(t) = -t\phi(t)$, we have:

$$
\frac{\partial f_\alpha}{\partial t} = \frac{\phi(t)}{\alpha(t)^2}(\alpha(t)^2 - t \cdot \alpha(t) - \alpha'(t)).
$$
Thus, the sign of the partial derivative is determined by \( g_\alpha(t) := \alpha(t)^2 - t \cdot \alpha(t) - \alpha'(t) \). We now consider four cases:

1. \( \alpha(t) = t + 1/t: \) In this case, \( g_\alpha(t) = 2/t^2 > 0. \) Therefore, \( f_\alpha(t) \) is strictly increasing in \( t \); together with \( f_\alpha(0) = -\frac{1}{2} \) and \( \lim_{t \to \infty} f_\alpha(t) = 0 \), it follows that \( f_\alpha(t) < 0 \) for all \( t \geq 0 \).

2. \( \alpha(t) = t + \lambda_0: \) In this case, \( g_\alpha(t) = \lambda_0 t + \lambda_0^2 - 1 \) is linear in \( t \). Set \( d := (1 - \lambda_0^2)/\lambda_0 > 0 \), and it follows that \( g_\alpha(t) < 0 \) for \( 0 \leq t < d \) and \( g_\alpha(t) > 0 \) for \( t > d \). Therefore, \( f_\alpha(t) \) is decreasing in \( t \) over \([0, d]\) and increasing in \( t \) over \([d, \infty)\); the endpoint conditions imply that \( f_\alpha(t) \leq 0 \) for all \( t \geq 0 \) with equality only at \( t = 0 \).

3. \( \alpha(t) = t: \) In this case, \( g_\alpha(t) = -1 \) so \( f_\alpha(t) \) is strictly decreasing in \( t \). Now \( \lim_{t \to 0} f_\alpha(t) = \infty \). Therefore, \( f_\alpha(t) > 0 \) for all \( t \geq 0 \).

4. \( \alpha(t) = \lambda_0^2 \cdot t + \lambda_0: \) In this case, \( g_\alpha(t) \) is quadratic in \( t \) with a zero constant term. Set \( d := \frac{2\lambda_0^2 - 1}{\lambda_0(1-\lambda_0)} > 0 \) and an easy calculation shows that \( g_\alpha(t) > 0 \) for \( 0 \leq t < d \) and \( g_\alpha(t) \leq 0 \) for \( t > d \). An analogous argument implies that \( f_\alpha(t) \geq 0 \) for all \( t \geq 0 \) with equality only at \( t = 0 \).

\[ \square \]

We now show two interesting properties of the tail probability function. These seem to be new as far as we know.

**Lemma 2.6.3.** The function \( Q(t)^{1/t^2} \) is increasing in \( t \) for \( t \geq 0 \). For every fixed \( 0 \leq a \leq 1 \), the function \( Q(at)/Q(t) \) is increasing in \( t \) for \( t \geq 0 \).

**Proof.** We use the basic identities \( (\ln Q(t))' = -\lambda(t) \) and \( \lambda'(t) = \lambda(t)^2 - t\lambda(t) \).

For the first property, it suffices to show that the function \( f(t) := \frac{1}{t^2} \cdot \ln Q(t) \) is increasing in \( t \) for \( t \geq 0 \). We have:

\[
\frac{df}{dt} = -\frac{t\lambda(t) + 2\ln(Q(t))}{t^3}.
\]

Let \( u(t) := t\lambda(t) + 2\ln(Q(t)) \) and observe that \( u'(t) = \lambda(t)(t\cdot\lambda(t) - t^2 - 1) < 0 \) by Lemma 2.6.2. Thus, \( f'(t) > 0 \) and \( f(t) \) is increasing in \( t \).

For the second property, it suffices to show that, for each fixed \( 0 \leq a \leq 1 \), the function \( g(t, a) := \ln Q(at) - \ln Q(t) \) is increasing in \( t \) for \( t \geq 0 \). We have:

\[
\frac{\partial g}{\partial t} = \lambda(t) - a \cdot \lambda(at).
\]

At \( t = 0 \), the right side equals 0, and we will show that \( \lambda(t) > a \cdot \lambda(at) \) for \( t > 0 \). This would imply the desired property that \( g(t, a) \) is increasing in \( t \). Multiplying both sides by \( t \), we need to show that \( t \cdot \lambda(t) > at \cdot \lambda(at) \), i.e., that the function \( h(x) := x \cdot \lambda(x) \) is an increasing function of \( x \) for \( x \geq 0 \). This holds because \( h'(x) = \lambda(x)(1 - x^2 + x\lambda(x)) > 0 \) by Lemma 2.6.2. \[ \square \]
We are ready to prove Proposition 2.1.4.

**Proposition 2.6.4** (Proposition 2.1.4 restated). Let $t, \delta \geq 0$. Set $\eta := \rho + \delta \sqrt{1 - \rho^2}$ and $\lambda_0 := \sqrt{\frac{2}{\pi}}$. Then:

(a) $\frac{e^{-t^2/2}}{t + \lambda_0} \preceq Q(t) \preceq \frac{e^{-t^2/2}}{t + 1/\lambda_0} \preceq e^{-t^2/2}$;
(b) $\frac{Q(t)^{\delta^2}}{\delta t + \lambda_0} \preceq Q(\delta t) \preceq Q(t)^{\delta^2} (t + \lambda_0)^2$;
(c) $L(t, \eta; \rho) \geq Q(t) Q(\delta t)$; and
(d) $Q(t) \leq Q(\delta t) \leq Q(t)^{\delta^2}$, if $\delta \leq 1$.

**Proof.** Substituting the definition of $\lambda(t)$ in Lemma 2.6.2 and simplifying the expression, we obtain (a). Applying these bounds appropriately on both sides of (b) proves that inequality as well.

Next, let $(X, Y) \sim \text{BGS}(\rho)$ so that $L(t, \eta; \rho) = \Pr[X > t, Y > \eta t]$. When $\rho = 1$, we have $X = Y$ with probability 1 so $L(t, \eta; \rho) = Q(t)$, implying (c) trivially. Henceforth, we let $\rho < 1$.

Now $Y = \rho X + \sqrt{1 - \rho^2} Z$ where $Z \sim \mathcal{N}(0, 1)$ is independent of $(X, Y)$. Observe:

\[
\begin{align*}
\Pr[X > t, Y > \eta t] &= \Pr[X > t, \rho X + \sqrt{1 - \rho^2} Z > \eta t] \\
&\geq \Pr[X > t, \rho t + \sqrt{1 - \rho^2} Z > \eta t] \\
&= \Pr[X > t, Z > \delta t] \quad \text{(valid, because $\rho < 1$)} \\
&= Q(t) \cdot Q(\delta t),
\end{align*}
\]

proving (c). For the last inequality, because $\delta \leq 1$, we have that $Q(t) \leq Q(\delta t)$, and the latter can be bounded from above using the first property in Lemma 2.6.3, which implies (d). \qed

### 2.7 Non-Asymptotic Bounds on Correlated Binomials

We let $h(\cdot)$ denote the binary entropy function. We start by stating the following two basic facts.

**Fact 2.7.1.** Stirling’s approximation of the factorial implies that for every integers $0 < \ell < m$, we have that

\[
\binom{m}{\ell} = \Theta\left(\sqrt{\frac{m}{\ell \cdot (m - \ell)}}\right) \cdot 2^{-m \cdot h(\frac{\ell}{m})}.
\]

**Fact 2.7.2** (Taylor approximation of binary entropy function). For every $x \in [0, 1]$, we have that

\[
h(1/2 - x/2) = 1 - \frac{1}{2 \ln 2} x^2 - O(x^4).
\]

We now prove Proposition 2.4.2.
Proof of Proposition 2.4.2. We have that

\[
\begin{align*}
\Pr_{X \in \mathbb{R}^n}_{\text{wt}(X) \leq n/2 - u\sqrt{n}/2} &= \sum_{i=0}^{n/2-u\sqrt{n}/2} \binom{n}{i} \cdot 2^{-n} \\
&\leq n \cdot 2^{-n} \cdot \binom{n}{n/2 - u\sqrt{n}/2} \\
&= n \cdot 2^{-n} \cdot \Theta\left(\sqrt{\frac{n}{n/2 - u\sqrt{n}/2}} \cdot \frac{n}{n/2 + u\sqrt{n}/2}\right) \cdot 2^{n \cdot h\left(\frac{n/2-u\sqrt{n}}{n}\right)} \\
&\leq O(n) \cdot 2^{-n} \cdot 2^{n \cdot \left(1 - \frac{u^2}{2\ln 2}\right)} \\
&= O(n) \cdot e^{-\frac{u^2}{2}} \\
&\leq O(n^2) \cdot Q(u),
\end{align*}
\]

where (A) follows from Fact 2.7.1, (B) from Fact 2.7.2, and (C) from proposition 2.1.4 (a). Since the distribution of \(\text{wt}(X)\) is symmetric around \(n/2\), the other case follows as well. \(\square\)

We point out that in the statement of Proposition 2.4.2 we made no effort to optimize the multiplicative function of \(n\) as that would not be consequential for our purposes. Recall that \(r := n/2 - t\sqrt{n}/2\).

Proposition 2.7.3. For any \(k = o(\sqrt{n})\), we have that

\[
\Pr_{X \in \mathbb{R}^n}_{\text{wt}(X) \leq r} \geq \frac{1}{\Theta(\sqrt{n})} \cdot Q(t).
\]
Proof. We have that
\[
\Pr_{X \in \{0,1\}^n} \{ \text{wt}(X) \leq r \} = \sum_{i=0}^{r} \binom{n}{i} \cdot 2^{-n} \\
\geq 2^{-n} \binom{n}{r} \\
= 2^{-n} \cdot \Theta \left( \sqrt{\frac{n}{r \cdot (n-r)}} \right) \cdot 2^{n-h \left( \frac{n/2-\sqrt{n/2}}{n} \right)} \\
\geq 2^{-n} \cdot \frac{1}{\Theta(\sqrt{n})} \cdot 2^{n-h \left( \frac{n/2-\sqrt{n/2}}{n} \right)} \\
= 2^{-n} \cdot \frac{1}{\Theta(\sqrt{n})} \cdot 2^n \left( 1 - \frac{\ell^2}{2m^2} - o \left( \frac{\ell^4}{m^4} \right) \right) \\
= \frac{1}{\Theta(\sqrt{n})} \cdot e^{-\ell^2 \frac{2}{n}} \\
\geq \frac{1}{\Theta(\sqrt{n})} \cdot Q(t),
\]
where the second equality follows from Fact 2.7.1, the third equality follows from Fact 2.7.2, the fourth equality uses the assumption that \( k = o(\sqrt{n}) \) and the fact that \( t = \Theta(\sqrt{k}) \), and the last inequality follows from Proposition 2.1.4(a). \( \square \)

Lemma 2.7.4. Fix \( \epsilon \in (0, 0.5] \). For positive every \( \alpha \) such that \( \alpha^3 \cdot m = o_m(1) \), we have that
\[
\Pr[\text{Bin}(m, \epsilon) = (\epsilon + \alpha) \cdot m] \geq \Theta \left( \frac{1}{\sqrt{m}} \right) \cdot e^{-\frac{m \cdot \alpha^2}{2(1-\epsilon)^3}},
\]
and similarly,
\[
\Pr[\text{Bin}(m, \epsilon) = (\epsilon - \alpha) \cdot m] \geq \Theta \left( \frac{1}{\sqrt{m}} \right) \cdot e^{-\frac{m \cdot \alpha^2}{2(1-\epsilon)}},
\]
Proof. Stirling’s approximation of the factorial implies that for every integers \( 0 < \ell < m \), we have that
\[
\binom{m}{\ell} = \Theta \left( \sqrt{\frac{m}{\ell \cdot (m-\ell)}} \right) \cdot 2^{m-h \left( \frac{\ell}{m} \right)}. \quad (2.19)
\]
Applying Equation (2.19) with \( \ell \triangleq (\epsilon + \alpha) \cdot m \), we get that
\[
\binom{m}{(\epsilon + \alpha) \cdot m} \geq \Theta \left( \frac{1}{\sqrt{m}} \right) \cdot 2^{m-h(\epsilon + \alpha)}.
\]
Thus,

\[
\Pr[\text{Bin}(m, \epsilon) = (\epsilon + \alpha) \cdot m] = \left(\frac{m}{\epsilon + \alpha} \cdot m\right) \cdot \epsilon^{(\epsilon+\alpha)\cdot m} \cdot (1 - \epsilon)^{m-(\epsilon+\alpha)\cdot m}
\]

\[
\geq \Theta\left(\frac{1}{\sqrt{m}}\right) \cdot 2^{m \cdot h(\epsilon + \alpha)} \cdot 2^{m \cdot (-h(\epsilon + \alpha) + \alpha \cdot \log \left(\frac{\alpha}{1-\epsilon}\right))}
\]

\[
= \Theta\left(\frac{1}{\sqrt{m}}\right) \cdot 2^{m \cdot (h(\epsilon + \alpha) - h(\epsilon) + \alpha \cdot \log \left(\frac{\epsilon}{1-\epsilon}\right))}.
\]

Note that for every \( \epsilon > 0 \),

\[
h'(\epsilon) = -\log \left(\frac{\epsilon}{1-\epsilon}\right),
\]

and

\[
h''(\epsilon) = -\frac{1}{\ln 2 \cdot \epsilon \cdot (1-\epsilon)}.
\]

Taylor expanding \( h(\epsilon + \alpha) \) around \( \epsilon > 0 \), we get that

\[
h(\epsilon + \alpha) = h(\epsilon) + h'(\epsilon) \cdot \alpha + \frac{h''(\epsilon) \cdot \alpha^2}{2} \pm O_\epsilon(\alpha^3)
\]

\[
= h(\epsilon) - \alpha \cdot \log \left(\frac{\epsilon}{1-\epsilon}\right) - \frac{\alpha^2}{2 \cdot \ln 2 \cdot \epsilon \cdot (1-\epsilon)} \pm O_\epsilon(\alpha^3).
\]

Thus, we get that

\[
\Pr[\text{Bin}(m, \epsilon) = (\epsilon + \alpha) \cdot m] \geq \Theta\left(\frac{1}{\sqrt{m}}\right) \cdot 2^{m \cdot \left(-\frac{\alpha^2}{2 \cdot \ln 2 \cdot \epsilon \cdot (1-\epsilon)} \pm O_\epsilon(\alpha^3)\right)}
\]

\[
= \Theta\left(\frac{1}{\sqrt{m}}\right) \cdot e^{-\frac{m \cdot \alpha^2}{2 \cdot \ln 2 \cdot (1-\epsilon)}},
\]

where the last equality uses the given assumption that \( \alpha^3 \cdot m = o_m(1) \). The proof of the second part of the lemma follows along the same lines with \( \alpha \) being replaced by \(-\alpha\). \( \square \)

**Proposition 2.7.5.** Fix \( \epsilon \in (0, 0.5] \). For every \( n = \omega(k^3) \), we have that

\[
\Pr_{(X,Y) \sim \text{DSBS}(1-2\epsilon)^{\otimes n}}[Y \in \text{Ball}(0, r') \mid X \in \text{Ball}(0, r)] \geq \Theta\left(\frac{1}{n^{1.5}}\right) \cdot 2^{-\delta^2 k},
\]

where \( \text{Ball}(0, r) \) denotes the Hamming ball of radius \( r \) centered around the all-zeros vector.

**Proof.** We start by showing that if \( A \sim \text{Bin}(\epsilon, n/2 + t\sqrt{n}/2) \) and \( B \sim \text{Bin}(\epsilon, n/2 - t\sqrt{n}/2) \) are independent random variables, then

\[
\Pr[A \leq B + r' - r] \geq \Theta\left(\frac{1}{n}\right) \cdot e^{-\frac{\delta^2 k}{2}}.
\]

(2.20)
To prove Equation (2.20), note that

\[
\Pr[A \leq B + r' - r] \geq \Pr[A = \frac{\epsilon \cdot n + 0.5\eta t \sqrt{n}}{2}, B = \frac{\epsilon \cdot n - 0.5\eta t \sqrt{n}}{2}]
= \Pr[A = \frac{\epsilon \cdot n + 0.5\eta t \sqrt{n}}{2}] \cdot \Pr[B = \frac{\epsilon \cdot n - 0.5\eta t \sqrt{n}}{2}].
\]

Applying Lemma 2.7.4 with \(m = n/2 + t \sqrt{n}/2\) and \(\alpha = \frac{-\epsilon + 0.5\eta t}{\sqrt{n+t}}\), we get that

\[
\Pr[A = \frac{\epsilon \cdot n + 0.5\eta t \sqrt{n}}{2}] \geq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot e^{-\frac{(\epsilon-0.5\eta)^2 t^2}{4 + (1-\epsilon)}}.
\]

Similarly, applying Lemma 2.7.4 with \(m = n/2 - t \sqrt{n}/2\) and \(\alpha = \frac{\epsilon-0.5\eta t}{\sqrt{n-t}}\), we get that

\[
\Pr[B = \frac{\epsilon \cdot n - 0.5\eta t \sqrt{n}}{2}] \geq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot e^{-\frac{(\epsilon-0.5\eta)^2 t^2}{4 + (1-\epsilon)}}.
\]

Note that when applying Lemma 2.7.4, we have used the assumption that \(n = \omega(k^3)\) and the fact that \(t = \Theta(\sqrt{k})\). Thus, we get that

\[
\Pr[A \leq B + r' - r] \geq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot e^{-\frac{t^2}{2}} \cdot e^{-\frac{(\epsilon-0.5\eta)^2 t^2}{4 + (1-\epsilon)}} \cdot \frac{1}{n}.
\]

where the last equality follows from the fact that \(n = \omega(t^4)\), which in particular follows from the assumption that \(n = \Omega(k^3)\) and the fact that \(t = \Theta(\sqrt{k})\). Equation (2.20) now implies that

\[
\Pr_{(X,Y) \sim \text{DSBS}(1-2e)^\otimes n}[Y \in \text{Ball}(0, r')|\text{wt}(X) = r] \geq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-\delta^2 k},
\]

where \(\text{wt}(X)\) denotes the Hamming weight of \(X\). The statement of Proposition 2.7.5 now follows from the fact that

\[
\Pr_{X \in R(0,1)^n}[\text{wt}(X) = r | X \in \text{Ball}(0, r)] \geq \Theta\left(\frac{1}{\sqrt{n}}\right),
\]

which itself uses the fact that \(r \leq n/2\).

In order to prove Lemma 2.4.4, we will need the following propositions.

**Proposition 2.7.6.** Let \(t_2 \geq 0\) and \(a_{\max} \triangleq n \cdot (1 + \theta)/4 - (t + t_2) \cdot \sqrt{n}/4\). For every
\( a \in \{0, 1, \ldots, a_{\text{max}}\} \), let
\[
\psi(a) \triangleq \frac{\binom{n \cdot (1+\theta)/2}{a}}{\binom{n/2 - t_2 \sqrt{n}/2 - a}{n/2 - t_2 \sqrt{n}/2}}.
\]

Then, \( \psi(a) \) is monotonically increasing in \( a \).

**Proof.** Let \( a \in \{1, \ldots, a_{\text{max}}\} \). Then,
\[
\frac{\psi(a)}{\psi(a-1)} = \frac{(n/2 \cdot (1+\theta) + 1 - a) \cdot (n/2 - t_2 \sqrt{n}/2 + 1 - a)}{a \cdot (t_2 \cdot \sqrt{n}/2 - \theta \cdot n/2 + a)}.
\]

This implies that \( \psi(a) \geq \psi(a-1) \) if and only if
\[
a \leq \frac{(n/2 \cdot (1+\theta) + 1) \cdot (n/2 - t_2 \sqrt{n}/2 + 1)}{n + 2},
\]
which is satisfied by all \( a \in \{0, 1, \ldots, a_{\text{max}}\} \) (for large enough \( n \)). \( \square \)

**Proposition 2.7.7.** Assume that \( t = o(n^{1/4}) \), \( t_2 = o(n^{1/4}) \) and \( \theta \cdot t \cdot t_2 = o_n(1) \). Then,
\[
\psi(a_{\text{max}}) \leq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot e^{-t^2/2}.
\]

**Proof.** By Fact 2.7.1, we have that
\[
\left(\frac{n \cdot (1+\theta)/2}{a_{\text{max}}}\right) = \Theta\left(\sqrt{\frac{n \cdot (1+\theta)/2}{a_{\text{max}} \cdot (n \cdot (1+\theta)/2 - a_{\text{max}})}}\right) \cdot 2^{-n \frac{(1+\theta)}{2} h\left(\frac{a_{\text{max}}}{n \cdot (1+\theta)/2}\right)}
\]
\[
= \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-n \frac{(1+\theta)}{2} h\left(\frac{1}{2} - \frac{(t_2 + t)^2}{2(1+\theta) \sqrt{n}}\right)}
\]
(2.21)

(where the second equality uses the assumptions that \( t = o(\sqrt{n}) \), \( t_2 = o(\sqrt{n}) \) and \( \theta = o_n(1) \)),

\[
\left(\frac{n \cdot (1 - \theta)/2}{n/2 - t_2 \sqrt{n}/2 - a_{\text{max}}}\right)
\]
\[
= \Theta\left(\sqrt{\frac{n(1-\theta)/2}{(n/2 - t_2 \sqrt{n}/2 - a_{\text{max}}) (t_2 \sqrt{n}/2 - nth/2 + a_{\text{max}})}}\right) \cdot 2^{-n \frac{(1-\theta)}{2} h\left(\frac{n/2 - t_2 \sqrt{n}/2 - a_{\text{max}}}{n(1-\theta)/2}\right)}
\]
\[
= \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-n \frac{(1-\theta)}{2} h\left(\frac{1}{2} - \frac{(t_2 - t)^2}{2(1-\theta) \sqrt{n}}\right)}
\]
(2.22)
where the second equality uses the assumption that \( t_2 = o(\sqrt{n}) \). Combining Equation (2.21), Equation (2.22) and Equation (2.23), we get that

\[
\psi(a_{\text{max}}) = \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-n} \left(\frac{(1+\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t+t_2)}{2 \cdot (1+\theta) \cdot \sqrt{n}}\right) + \frac{(1-\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t_2-t)}{2 \cdot (1-\theta) \cdot \sqrt{n}}\right) - h\left(\frac{1}{2} - \frac{t_2}{2 \cdot \sqrt{n}}\right)\right). \tag{2.24}
\]

By Fact 2.7.2, we have that

\[
\frac{(1+\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t+t_2)}{2 \cdot (1+\theta) \cdot \sqrt{n}}\right) + \frac{(1-\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t_2-t)}{2 \cdot (1-\theta) \cdot \sqrt{n}}\right) = \frac{(1+\theta)}{2} \left(1 - \frac{(t+t_2)^2}{2 \ln 2 \cdot (1+\theta)^2 n^2}\right) + \frac{(1-\theta)}{2} \left(1 - \frac{(t_2-t)^2}{2 \ln 2 \cdot (1-\theta)^2 n^2}\right) - \frac{t_2^2}{2 \ln 2 \cdot n} + O\left(\frac{(t+t_2)^4}{n^2}\right), \tag{2.25}
\]

where the second equality above uses the fact that \( \theta \leq 1 \). Plugging Equation (2.25) back in Equation (2.24), we get that

\[
\psi(a_{\text{max}}) = \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-n} \left(\frac{(1+\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t+t_2)}{2 \cdot (1+\theta) \cdot \sqrt{n}}\right) + \frac{(1-\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t_2-t)}{2 \cdot (1-\theta) \cdot \sqrt{n}}\right) - h\left(\frac{1}{2} - \frac{t_2}{2 \cdot \sqrt{n}}\right)\right) + O\left(\frac{(t+t_2)^4}{n^2}\right)
\]

\[
\leq \Theta\left(\frac{1}{\sqrt{n}}\right) \cdot 2^{-n} \left(\frac{(1+\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t+t_2)}{2 \cdot (1+\theta) \cdot \sqrt{n}}\right) + \frac{(1-\theta)}{2} \cdot h\left(\frac{1}{2} - \frac{(t_2-t)}{2 \cdot (1-\theta) \cdot \sqrt{n}}\right) - h\left(\frac{1}{2} - \frac{t_2}{2 \cdot \sqrt{n}}\right)\right) + O\left(\frac{(t+t_2)^4}{n^2}\right)
\]

where the inequality uses the assumptions that \( t = o(n^{1/4}) \), \( t_2 = o(n^{1/4}) \) and \( \theta \cdot t \cdot t_2 = o_n(1) \). \( \square \)

### 2.8 Correlated Randomness Generation

We first recall that Canonne et al. \[CGMS14\] – using the converse bound of \[BM11\] – showed that for any \( \epsilon > 0 \), if Alice and Bob are given access to i.i.d. samples from \( DSBS(1-2\epsilon) \), then, perfectly agreeing on \( k \) random bits requires \( \Omega(k) \) bits of communication even in the two-way model. They also raised the following intriguing question: “What if their goal is only to generate more correlated bits than they start with? What is possible here and what
are the limits?"

We partially answer this question and show that for any \( \epsilon > 0 \) and \( \epsilon' \gg \epsilon \cdot \log(1/\epsilon) \), if Alice and Bob are given access to i.i.d. samples from \( \text{DSBS}(1-2\epsilon') \), then, generating \( k \) random samples from \( \text{DSBS}(1-2\epsilon) \) requires \( \Omega_{\epsilon',\epsilon}(k) \) bits of communication.

**Definition 2.8.1 (Correlated Randomness Generation).** In the CorrelatedRandomness\(_{\gamma,\epsilon',\alpha,k} \) problem, Alice and Bob are given access to i.i.d. samples from a known source \( \mu \). Their goal is to for Alice to output \( w_A \in \{0,1\}^k \) and for Bob to output \( w_B \in \{0,1\}^k \) that satisfy the following properties: (i) \( \Pr[\Delta(w_A, w_B) \leq \epsilon'k] \geq \gamma \); (ii) \( H_\infty(w_A) \geq \alpha \cdot k \); and (iii) \( H_\infty(w_B) \geq \alpha \cdot k \).

We point out that one can alternatively define Correlated Randomness Generation in terms of coming close, say in total variation distance, to the distribution \( \text{DSBS}(1-2(\epsilon'-\delta)) \). The results in this section apply to this variant as well. This is because of the next lemma which can be proved by a simple Chernoff bound and which says that if Alice and Bob are given access to i.i.d. samples from \( \text{DSBS}(1-2\epsilon) \), then they can generate two length-\( k \) binary strings that lie in a Hamming ball of radius \( \approx \epsilon' \cdot k \) with high probability.

**Lemma 2.8.2.** Fix \( 0 < \delta < \epsilon' \) and let \( \text{DSBS}(1-2(\epsilon'-\delta)) \) be the source. Then, there is a non-interactive protocol solving CorrelatedRandomness\(_{\gamma,\epsilon',\alpha,k} \) with \( \gamma = 1 - \exp(-((\epsilon'-\delta)^2 \cdot k) \) and \( \alpha = 1 \).

We are now ready to state the main result.

**Theorem 2.8.3 (Interactive Correlated Randomness Generation).** Any interactive protocol solving CorrelatedRandomness\(_{\gamma,\epsilon',\alpha,k} \) for the source \( \text{DSBS}(1-2(\epsilon'-\delta)) \) with \( h(\epsilon') \leq 4 \cdot \epsilon \cdot (1-\epsilon) \cdot \alpha/(1+\Omega(1)) \) should communicate at least \( \Omega(\epsilon \cdot \alpha \cdot k) - O(\log(1/\gamma)) \) bits.

Theorem 2.8.4 says that non-interactively generating two strings with min-entropy \( k \) and that lie in a Hamming ball of radius \( \approx \epsilon' \cdot k \) cannot be done with success probability \( 2^{-\alpha(k)} \) when Alice and Bob are given access to i.i.d. samples from \( \text{DSBS}(1-2\epsilon) \) with \( \epsilon = \omega(\epsilon' \cdot \log(1/\epsilon')) \).

**Theorem 2.8.4 (Non-Interactive Correlated Randomness Generation).** There is no non-interactive protocol solving CorrelatedRandomness\(_{\gamma,\epsilon',\alpha,k} \) for the source \( \text{DSBS}(1-2\epsilon) \) with \( h(\epsilon') \leq 4 \cdot \epsilon \cdot (1-\epsilon) \cdot \alpha \) and \( \gamma > 2^{-\nu k} \) where

\[
\nu = \alpha \cdot \left[ \frac{\sqrt{1 - h(\epsilon')/\alpha} - (1 - 2\epsilon)}{4 \cdot \epsilon \cdot (1 - \epsilon)} \right]^2.
\]

Consequently, whenever \( h(\epsilon') \leq 4 \cdot \epsilon \cdot (1-\epsilon) \cdot \alpha/(1+\Omega(1)) \), there is no non-interactive protocol solving CorrelatedRandomness\(_{\gamma,\epsilon',\alpha,k} \) given i.i.d. access to \( \text{DSBS}(1-2\epsilon) \) with \( \gamma > 2^{-\Omega(\epsilon \cdot \alpha \cdot k)} \).

We point out that getting the tight bounds in Theorem 2.8.3 and Theorem 2.8.4 remains a very interesting open question. In order to prove Theorem 2.8.3 and Theorem 2.8.4, we next introduce a “list” version of Common Randomness which is implicit in several of the known converse results for Common Randomness Generation.
Definition 2.8.5 (List Common Randomness Generation). In the ListCommonRandomness\(_k\) problem, Alice and Bob are given access to i.i.d. samples from a known distribution \(\mu\) over pairs of random variables. Their goal is for Alice to output an element \(w_A\) and for Bob to output a list \(L_B\) (over the same universe) such that (i) \(\Pr[w_A \in L_B] \geq \gamma\); (ii) \(H_{\min}(w_A) \geq k\); and (iii) \(|L_B| \leq b\).

We prove the following converse results for List Common Randomness Generation both in the non-interactive and two-way communication models:

Theorem 2.8.6 (Non-Interactive List Common Randomness Generation). There is no non-interactive protocol solving ListCommonRandomness\(_k\) for the source DSBS(1 - 2\(\epsilon\)) with \((\log b)/k \leq 4 \cdot \epsilon \cdot (1 - \epsilon)\) and with \(\gamma > 2^{-\nu k}\) where
\[
\nu = \frac{[\sqrt{1 - (\log b)/k} - (1 - 2\epsilon)]^2}{4 \cdot \epsilon \cdot (1 - \epsilon)}.
\]
Consequently, whenever \((\log b)/k \leq 4 \cdot \epsilon \cdot (1 - \epsilon)/(1 + \Omega(1))\), there is no non-interactive protocol solving ListCommonRandomness\(_k\) with \(\gamma > 2^{-\Omega(\epsilon^{-k})}\).

Proof. The proof is very similar to that of the converse result of [GR16]. Let \(\Pi\) be a protocol solving ListCommonRandomness\(_{\gamma,b}\). Let \(X\) be Alice’s input and \(w_A \triangleq f(X)\) be her output, and let \(Y\) be Bob’s input and \(L_B \triangleq (g_1(Y), g_2(Y), \ldots, g_b(Y))\) be his output. Here, \((X,Y) \sim DSBS(1 - 2\epsilon)^\otimes n\), and \(f, g_1, g_2, \ldots, g_b\) are functions mapping \(\{0,1\}^n\) to \(\{0,1\}^k\). For every \(y \in \{0,1\}^n\) and \(z \in \{0,1\}^k\), denote \(\beta(z|y) \triangleq \Pr[f(X) = z|Y = y]\). The success probability of the protocol \(\Pi\) is given by
\[
\Pr[w_A \in L_B] = \Pr[f(X) \in \{g_1(Y), g_2(Y), \ldots, g_b(Y)\}]
= \mathbb{E}_y[\Pr[f(X) \in L_B(y) \mid Y = y]]
= \mathbb{E}_y[\sum_{z \in L_B(y)} \beta(z|y)]
\leq \mathbb{E}_y[(\sum_{z \in L_B(y)} \beta(z|y)^q)^{1/q}] \cdot b^{1-1/q}
\leq \mathbb{E}_y[(\sum_{z} \beta(z|y)^q)^{1/q}] \cdot b^{1-1/q}
\leq (\mathbb{E}_y[\sum_{z} \beta(z|y)^q])^{1/q} \cdot b^{1-1/q}
= (\sum_{z} \mathbb{E}_y[\beta(z|y)^q])^{1/q} \cdot b^{1-1/q},
\]
where the first inequality follows from Holder’s inequality and the last inequality follows from the fact that the function \(x \mapsto x^{1/q}\) for non-negative \(x\) is concave for every \(q \geq 1\). Consider the function \(h_z: \{0,1\}^n \to \{0,1\}\) given by \(h_z(X) = 1[f(X) = z]\) for all \(X \in \{0,1\}^n\).
Hypercontractivity then implies that

\[
\mathbb{E}_y[\beta(y|y)^q]^{1/q} = \mathbb{E}_y[\mathbb{E}[h_z(X) \mid Y = y]^q]
\]
\[
= ||\mathbb{E}[h_z(X) \mid Y]^q||_p^q
\]
\[
\leq ||h_z||_p^q
\]
\[
= (\mathbb{E}_x h_z(x))^{q/p}
\]
\[
= \Pr[f(x) = z]^{q/p}.
\]

Thus, the success probability of \(\Pi\) satisfies

\[
\Pr[w_A \in L_B] \leq (\sum_z \Pr[f(X) = z]^{q/p})^{1/q} \cdot b^{1-1/q}
\]
\[
= (\sum_z \Pr[f(X) = z]^{q/p-1} \cdot \Pr[f(X) = z])^{1/q} \cdot b^{1-1/q}
\]
\[
\leq (2^{-k(\frac{q}{p}-1)} \cdot \sum_z \Pr[f(X) = z])^{1/q} \cdot b^{1-1/q}
\]
\[
= 2^{-k\cdot(\frac{q}{p}-1)} \cdot b^{1-1/q},
\]

where the inequality above follows from the fact that \(w_A\) has min-entropy at least \(k\) bits. Setting \(p = 1 + (1 - 2 \cdot \epsilon)^2 \cdot \delta\) and \(q = 1 + \delta\) and optimizing for \(\delta\), we get that

\[
\gamma \leq 2^{-k(\frac{\sqrt{s^2 + \sqrt{(1-(\log b)/k)^2}}}{1-s})},
\]

where \(s = (1 - 2\epsilon)^2\) is the Strong Data Processing Constant of the DSBS\((1-2\epsilon)\) source, and where the above bound holds assuming that \((\log b)/k \leq 1 - s\). The theorem statement now follows.

We point out that Theorem 2.8.6 implies a lower bound on the 1-way communication complexity of List Common Randomness Generation (by essentially increasing the list size by a factor of \(2^c\) where \(c\) is the communication from Alice to Bob). It turns out that, by adapting a reduction of [CGMS14], one can also use Theorem 2.8.6 to get a lower bound on the interactive communication complexity of List Common Randomness Generation, which we state next.

**Theorem 2.8.7** (Interactive List Common Randomness Generation). Let DSBS\((1-2\epsilon)\) be the source. Then, any interactive protocol solving ListCommonRandomness\(_k\) with \((\log b)/k \leq 4 \cdot \epsilon \cdot (1 - \epsilon)\) should communicate at least

\[
k \cdot \frac{[\sqrt{1 - (\log b)/k} - (1 - 2\epsilon)]^2}{8 \cdot \epsilon \cdot (1 - \epsilon)} - \frac{3}{2} \log(1/\gamma) - O(1) \quad \text{bits.}
\]

Consequently, whenever \((\log b)/k \leq 4 \cdot \epsilon \cdot (1 - \epsilon)/(1 + \Omega(1))\), any interactive protocol solving ListCommonRandomness\(_k\) should communicate at least \(\Omega(\epsilon \cdot k) - O(\log 1/\gamma)\) bits.
Proof. The proof will combine Theorem 2.8.6 with the approach of [CGMS14] for getting lower bounds on interactive Common Randomness Generation using lower bounds on non-interactive Common Randomness Generation.

Let \( \Pi \) be an interactive protocol solving \( \text{ListCommonRandomness}^k \) with \( (\log b)/k \leq (1 - s)/(1 + \Omega(1)) \). Let \( X \) denote Alice’s input and \( Y \) denote Bob’s input. Consider now the non-interactive protocol \( \Pi' \) where on input pair \((X, Y)\):

- Alice samples \( Y' \) from the conditional distribution of \( \mu \) given \( X \), and she outputs the element that she would have output in the execution of \( \Pi \) on \((X, Y')\).

- Bob samples \( X' \) from the conditional distribution of \( \mu \) given \( Y \), and he outputs the list that he would have output in the execution of \( \Pi \) on \((X', Y)\).

Note that the non-interactive protocol \( \Pi' \) satisfies the property that the min-entropy of Alice’s output is at least \( k \) (since it is exactly equal to the min-entropy of Alice’s output under \( \Pi \)). We next show that the success probability of the protocol \( \Pi' \) is at least \( \Omega(\gamma^3 \cdot 2^{-2c}) \) where \( c \) is the two-way communication complexity of \( \Pi \). Using Theorem 2.8.6, this would imply that

\[
c \geq k \cdot \frac{[\sqrt{1 - (\log b)/k} - (1 - 2\epsilon)]^2}{8 \cdot \epsilon \cdot (1 - \epsilon)} - \frac{3}{2} \log(1/\gamma) - O(1),
\]

which implies the desired statement. We now lower-bound the success probability of \( \Pi' \). Let \( P_X(t) \) denote the probability over \( Y' \) conditioned on \( X \) that \( \Pi(X, Y') \) is equal to the transcript \( t \). Similarly, let \( Q_Y(t) \) denote the probability over \( X' \) conditioned on \( Y \) that \( \Pi(X', Y) \) is equal to the transcript \( t \). Let \( G \) be the set of all input pairs \((X, Y)\) such that, in the execution of \( \Pi(X, Y) \), Alice’s output element belongs to Bob’s output list. Then, the success probability of \( \Pi \) is equal to

\[
\gamma = \sum_{(X, Y) \in G} \mu(X, Y).
\]

We say that a transcript \( t \) is unlikely for \( X \) if \( P_X(t) < (\gamma/4) \cdot 2^{-c} \). Similarly, we say that a transcript \( t \) is unlikely for \( Y \) if \( Q_Y(t) < (\gamma/4) \cdot 2^{-c} \). Let \( B \) be the set of all input-pairs \((X, Y)\) such that the transcript \( \Pi(X, Y) \) is either unlikely for \( X \) or unlikely for \( Y \). Note that

\[
\sum_{(X, Y): \Pi(X, Y) \text{ unlikely for } X} \mu(X, Y) = \sum_X \sum_{t \text{ unlikely for } X} \sum_{Y: \Pi(X, Y) = t} \mu(X, Y)
\]

\[
= \sum_X \mu(X) \cdot \sum_{t \text{ unlikely for } X} P_X(t)
\]

\[
< \sum_X \mu(X) \cdot \sum_{t \text{ unlikely for } X} \frac{\gamma}{4} \cdot 2^{-c}
\]

\[
< \frac{\gamma}{4}.
\]

(2.26)

54
An identical argument shows that
\[
\sum_{(X,Y): \Pi(X,Y) \text{ unlikely for } Y} \mu(X,Y) < \frac{\gamma}{4}. \tag{2.27}
\]
Combining Equation (2.26) and Equation (2.27), we get that
\[
\sum_{(X,Y)\in B} \mu(X,Y) < \frac{\gamma}{2}.
\]

The success probability of \(\Pi'\) can now be lower-bounded by
\[
\sum_{(X,Y)\in G} \mu(X,Y) \cdot P_X(\Pi(X,Y)) \cdot Q_Y(\Pi(X,Y)) \geq \sum_{(X,Y)\in G\setminus B} \mu(X,Y) \cdot \frac{\gamma^2}{16} \cdot 2^{-2-c}
\]
\[
= \frac{\gamma^2}{16} \cdot 2^{-2-c} \cdot \left( \sum_{(X,Y)\in G} \mu(X,Y) - \sum_{(X,Y)\in B} \mu(X,Y) \right)
\]
\[
\geq \frac{\gamma^3}{32} \cdot 2^{-2-c},
\]
as desired. \(\square\)

We note that Theorem 2.8.6 and Theorem 2.8.7 also hold with the same bounds when the source is \(\text{BGS}(1-2\epsilon)\) instead of \(\text{DSBS}(1-2\epsilon)\). We now show how Theorem 2.8.6 implies Theorem 2.8.4, and how Theorem 2.8.7 implies Theorem 2.8.3.

**Proof of Theorem 2.8.4.** Given a protocol \(\Pi\) for \(\text{CorrelatedRandomness}_{\gamma,\epsilon',\alpha,k}\), we give a protocol \(\Pi'\) for \(\text{ListCommonRandomness}_{\alpha\cdot k}\) with \(b \leq 2^h(\epsilon')k\) as follows:

- If \(w_A\) is the output of Alice under the protocol \(\Pi\), then she also outputs \(w_A\) under the protocol \(\Pi'\).

- If \(w_B\) is the output of Bob under the protocol \(\Pi\), then he outputs the list \(L_B \triangleq \text{Ball}(w_B, \epsilon' \cdot k)\) under the protocol \(\Pi'\).

Theorem 2.8.4 now follows from Theorem 2.8.6 and the fact that \(|\text{Ball}(w_a, \epsilon' \cdot k)| \leq 2^h(\epsilon')k\). \(\square\)

**Proof of Theorem 2.8.3.** The proof is identical to that of Theorem 2.8.4 except that we use Theorem 2.8.7 instead of Theorem 2.8.6. \(\square\)
2.9 LSH and Common Randomness

On a high-level, a good locality sensitive hashing scheme maps, with high probability, nearby points to the same key, and far apart points to different keys. On the other hand, a good common randomness generation scheme (say for DSBS in the zero-communication case) has to map “nearby” points to the same key (since Alice’s and Bob’s inputs are relatively close to each other and have to map to the same key with noticeable probability). Then, to ensure the min-entropy requirement of the common randomness scheme, intuitively it seems that one has to map far apart points to different keys. This suggests a (at least intuitive) connection between locality sensitive hashing and common randomness generation. Formalizing this intuition remains a very intriguing open question.

In the following, we observe that in a certain regime, LSH and common randomness generation seem indeed to be related.

An important parameter that governs the performance of an LSH hash family $\mathcal{H}$ is given by its $\bar{\rho}(\mathcal{H})$ parameter [IM98]. Let $0 \leq \alpha \leq 1$ and $c \geq 1$. Loosely speaking, if the hash family ensures that points at relative distance at most $\alpha$ collide with probability at least $p_1$ while points at relative distance at least $c\alpha$ collide with probability at most $p_2$, then $\bar{\rho}(\mathcal{H}, \alpha, c) \leq \log(1/p_1)/\log(1/p_2)$. Smaller values of $\bar{\rho}(\mathcal{H}, \alpha, c)$ can potentially lead to improvements in the data structure performance. For the Hamming cube $\{0, 1\}^d$, there is a trivial scheme $\mathcal{H}_0$ such that $\bar{\rho}(\mathcal{H}_0) \leq \log\left(\frac{1}{(1 - 2\epsilon)}\right)/\log\left(\frac{1}{(1 - \epsilon)}\right) \to 1/c$ as $\alpha \to 0$.

We show that the zero-communication common randomness schemes considered here and in previous works [BM11, GR16] imply an LSH scheme with a strictly better $\bar{\rho}$ parameter. This is perhaps not surprising since the best strategy for a universal scheme is to map close-by points to the same output in order to achieve high-agreement probability, but to ensure high entropy it must map far-away points to different outputs.

Recall that in the trivial scheme $\mathcal{H}_0$ the hash function just outputs the bit at a random coordinate in $[d]$. When the relative distance between the two points is $\epsilon$, this is tantamount to producing a single sample from DSBS$(1 - 2\epsilon)$. Thus the trivial LSH scheme is also a trivial common randomness scheme using one sample from DSBS$(1 - 2\epsilon)$. If we use $k$ samples, i.e., take $k$ independent hash functions, and use the trivial scheme we obtain an agreement $p_\rho := \left(\frac{1 + \rho}{2}\right)^k$. Let $f_0(\rho) = \log(1/p_\rho)/k = \log\left(\frac{2}{(1 + \rho)}\right)$. In contrast, if we use the mapping given by the common randomness scheme then for this hash family (call it $\mathcal{H}_1$), the analogous expression equals $f_{CR}(\rho) := (1 - \rho)/(1 + \rho) + O(\log(k)/k)$. For large $k$ we can ignore the lower order term. So let $f_{CR}(\rho) = (1 - \rho)/(1 + \rho)$. To show that $\bar{\rho}(\mathcal{H}_2)$ is better we need to show for $\rho > \rho'$ that $f_{CR}(\rho)/f_{CR}(\rho') \leq f(\rho)/f(\rho')$. That is, $f(\rho)/f_{CR}(\rho)$ is an increasing function in $[0, 1]$. This can be verified analytically. In fact, it is always strictly increasing so the bound for the CR scheme is strictly better than the trivial one.

2.10 Min-Entropy Properties

In this section, we prove the min-entropy guarantee on Alice’s output.
2.10.1 DSBS($\rho$) ([BM11], rephrased).

Given a linear code $\mathcal{C} \subseteq \{0,1\}^n$ of dimension $d$, let $H_1, H_2, \ldots, H_{2^n-d}$ denote the cosets of $\mathcal{C}$ with $H_1 = \mathcal{C}$. For each $H_i$, fix a representative $a_i \in H_i$ so that $H_i = \mathcal{C} + a_i$. To maximize agreement, $a_i$ should be have the smallest weight in $H_i$. Let $f : \{0,1\}^n \rightarrow \mathcal{C}$ be defined as $f(v) = v - a_i$ where $v \in H_i$ for some unique $i$. Then $f^{-1}(c) = \{c + a_i : i \in [2^{n-d}]\}$, and $|f^{-1}(c)| = 2^{n-d}$, for all $c \in \mathcal{C}$. This proves that Alice’s output is uniformly distributed.

2.10.2 BGS($\rho$)

**Notation.** Let $X \overset{d}{=} Y$ denote that the probability distributions of $X$ and $Y$ are identical.

Let $G$ be a finite abelian group and let $\mathcal{C} = \{v_g : g \in G\} \subseteq \mathbb{R}^n$ be indexed by elements of $G$. They key to obtaining the min-entropy guarantee is the following:

For every $h \in G$ that there is a unitary transform $U_h$ such that $U_h v_g = v_{g+h}$ for all $g \in G$.

Let $(X, Y) \sim \text{BGS}(\rho)^{\otimes n}$. Abusing notation we denote $X_g := \langle v_g, X \rangle$ and $Y_g := \langle v_g, Y \rangle$. We may assume that with probability 1, there are no duplicates among the $X_g$’s and among the $Y_g$’s. This is because $v_g \neq v_{g'}$ for every $g \neq g'$ so the event $X_g = X_{g'}$ or $Y_g = Y_{g'}$ has measure zero.

**Lemma 2.10.1.** Fix $s \in G$ and suppose $U$ is a unitary transformation $U$ on $\mathbb{R}^n$ such that $U v_g = v_{g+s}$ for all $g \in G$. Then $(X_g, Y_g) \overset{d}{=} (X_{g+s}, Y_{g+s}) : g \in G$.

**Proof.** $(UX, UY)$ is a linear transform of $(X, Y)$ so it has the multivariate Gaussian distribution. Its marginals are standard Gaussian and its covariance matrix is identical that of $(X, Y)$ by elliptical symmetry. Thus, $(UX, UY) \overset{d}{=} (X, Y)$. Because $U$ is unitary, $X_g = \langle (Uv_g), (UX) \rangle$ and $Y_g = \langle (Uv_g), (UY) \rangle$ for all $g \in G$. We have:

$$(X_g, Y_g) : g \in G \overset{d}{=} ((\langle Uv_g), X), \langle (Uv_g), Y \rangle) : g \in G = (X_{g+s}, Y_{g+s}) : g \in G),$$

where the last step uses the property of $U$. \hfill \Box

**Lemma 2.10.2.** Define event $E_h := \{X_h = \max\{X_g : \forall g \in G\}\}$ for all $h \in G$. Then $\Pr[E_h] = \frac{1}{|G|}$ for all $h \in G$.

**Proof.** We have $\sum_h \Pr[E_h] = 1$ because the $X_g$’s are distinct with probability 1. We only need to show that $\Pr[E_h] = \Pr[E_{h'}]$ for all $h \neq h'$. Apply Lemma 2.10.1 with $s := h' - h$. We obtain:

$$\Pr[E_h] = \Pr[X_h = \max\{X_g : \forall g \in G\}] = \Pr[X_{h'} = \max\{X_{g+h'-h} : \forall g \in G\}] = \Pr[E_{h'}] \hfill \Box$$

Now fix a subgroup $K$ of $G$. Note that the cosets of $K$ partition $G$. 

57
Lemma 2.10.3. For every $h \in G$, define the event:

$$E_h := \left\{ \left( Y_h = \max \{Y_g; \forall g \in K + h\} \right) \land \left( \exists \ell \in K + h: X_\ell = \max \{X_g; \forall g \in G\} \right) \right\}$$

Then $\Pr[E_h] = \frac{1}{|G|}$ for all $h \in G$.

Proof. Similar to Lemma 2.10.2, observe that the union of events $E_h$ for $h \in H$ covers the entire space. Moreover, because with probability one, the $X_g$'s are distinct and the $Y_g$'s are distinct, we have $\sum_h \Pr[E_h] = 1$. We show that $\Pr[E_h] = \Pr[E_{h'}]$ for all $h \neq h'$.

Note that $t \in K + h$ if and only if $t + h' - h \in K + h'$ for all $t \in G$. Applying Lemma 2.10.1 with $s := h' - h$, we obtain:

$$\Pr[E_h] = \Pr\left[ \left( Y_h = \max \{Y_g; \forall g \in K + h\} \right) \land \left( \exists \ell \in K + h: X_\ell = \max \{X_g; \forall g \in G\} \right) \right]$$

$$= \Pr\left[ \left( Y_{h'} = \max \{Y_{g+h-h'}; \forall g \in K + h\} \right) \land \left( \exists \ell \in K + h: X_{\ell+h-h'} = \max \{X_{g+h-h'}; \forall g \in G\} \right) \right]$$

$$= \Pr\left[ \left( Y_{h'} = \max \{Y_{g}; \forall g \in K + h'\} \right) \land \left( \exists \ell \in K + h': X_\ell = \max \{X_g; \forall g \in G\} \right) \right]$$

$$= \Pr[E_{h'}] \quad \square$$

We apply the above to the two schemes for $\text{BGS}(\rho)$.

Simple construction. Recall that $C$ consists of the $n$ standard basis vectors $\{e_i: i \in [n]\}$ in $\mathbb{R}^n$. Let us re-index $i$ so that $i \in \mathbb{Z}_n$ and set $G = \mathbb{Z}_n$ By symmetry, for each $j \in \mathbb{Z}_n$, there is an elementary unitary transform $U_j$ using a permutation matrix such that $U_j e_i = e_{i+j}$ for all $i \in \mathbb{Z}_n$. For every $m$ dividing $n$, there exists a subgroup $K \subseteq G$ of size $m$, e.g., generated by $n/m$ in $\mathbb{Z}_n$.

Tao’s construction. Recall that $C := \{v_a: a \in \mathbb{F}_p^d\} \subseteq V$, where $V$ is the vector space of functions from $\mathbb{F}_p$ to $C$ and where for every $a \in \mathbb{F}_p^d$, $v_a(x) = \frac{1}{\sqrt{p}} \cdot \omega^{a_1x_1 + \cdots + a_dx_d}$. Here $G$ is defined by the additive structure of $\mathbb{F}_p^d$. For each $a \in \mathbb{F}_p^d$, define the map $U_a$ on $V$ as $(U_a v)(x) = \sqrt{p} \cdot v_a(x) v(x) = \omega^{a_1x_1 + \cdots + a_dx_d} \cdot v(x)$ for all $v \in V$. This is a unitary transform because for each $x \in \mathbb{F}_p$, $v(x)$ just undergoes a phase shift. Moreover $(U_b v_a)(x) = v_{a+b}(x)$, as needed. For every $0 \leq j \leq d$, there exists a subgroup $K \subseteq G$ of size $p^j$, e.g. $K = \{a \in \mathbb{F}_p^d: a_1 = a_2 = \cdots = a_{d-j} = 0\}$.

2.11 Conclusion and Open Questions

The most important open question raised in this chapter is to obtain *computationally* efficient schemes for common randomness. In particular, is there a resource-efficient scheme that also has time complexity $\text{poly}(k)$? For our schemes, it not at all clear how to implement the decoding phase time-efficiently (either over $\mathbb{F}_2$ or in Euclidean space). In fact, even the
slightly sub-exponential time algorithm of [KS13] for decoding dual-BCH codes falls short of working for the error radii that are needed to achieve near-optimal agreement probability!

The sample complexity $n = o(k^4)$ of our explicit schemes is polynomial but still far from the non-explicit schemes with linear sample complexity arising from amortized common randomness (in the case where the agreement probability tends to 1). The Kabatjanskii-Levenstein bound (cf. [Tao13]) implies that no nearly-orthogonal families of vectors (including the one we used) will achieve a linear sample complexity in our setup. Can we rule out linear sample schemes altogether (e.g., for small agreement probabilities)? One challenge is that such a proof cannot solely rely on hypercontractivity because they “tensorize” and are thus oblivious to the number $n$ of used samples.

Our one-way scheme for general sources with maximal correlation $\rho$ is explicit but not sample-efficient because it uses a Central Limit Theorem to reduce the problem to $BGS(\rho)$. Moreover, the trade-off between communication and agreement is stated in terms of $\rho$, whereas the best known negative results are in terms of hypercontractivity. [AGKN13] give an example of a source separating its maximal correlation from its Strong Data Processing Constant, which is intimately related to its hypercontractive properties. Can such a source be used to prove that the trade-off stated in Theorem 2.1.2 is not tight for general sources?

A characterization of amortized correlated randomness would be interesting even for one-way communication as it would generalize the Strong Data Processing Constant.

Finally, our work shows that tools used in common randomness could also be useful for Locality Sensitive Hashing. Can one establish a formal connection between these two areas?
Chapter 3

Non-Interactive Simulation of Joint Distributions

3.1 Introduction & Related Work

Given a sequence of independent samples \((x_1, y_1), (x_2, y_2), \ldots\) from a joint distribution \(P\) on \(\mathcal{A} \times \mathcal{B}\) where Alice observes \(x_1, x_2, \ldots\) and Bob observes \(y_1, y_2, \ldots\), what is the largest correlation that they can extract if Alice applies some function to her observations and Bob applies some function to his? The continuous version of this question – where the extracted correlation is required to be in Gaussian form – was solved by Witsenhausen in 1975 who gave (roughly) a \(\text{poly}(|\mathcal{A}|, |\mathcal{B}|, \log(1/\delta))\)-time algorithm that estimates the best such correlation up to an additive \(\delta\) [Wit75]. When the target distribution is Gaussian, the best possible correlation that is attainable is exactly the well-known “maximal correlation coefficient” which was first introduced by Hirschfeld [Hir35] and Gebelein [Geb41] and then studied by Rényi [Rén59]. However, when the target distribution is not Gaussian, the best correlation is not well-understood and this is the question explored in this chapter. Specifically, we study the Boolean version of this question where the extracted correlation is required to be in the form of bits with fixed specified marginals. We give an algorithm that, given \(\delta > 0\), computes the best such correlation up to an additive \(\delta\).

Questions such as the above are well-studied in the information theory literature under the label of “Non-Interactive Simulation”. The roots of this exploration go back to classical works by Gács and Körner [GK73] and Wyner [Wyn75]. In this line of work, the problem is described by a source distribution \(P(X, Y)\) and a target distribution \(Q(U, V)\) and the goal is to determine the maximum rate at which samples of \(P\) can be converted into samples of \(Q\). (So the goal is to start with \(n\) samples from \(P\) and generate \(R \cdot n\) samples from \(Q\), for the largest possible \(R\).) Gács and Körner considered the special case where \(Q\) required the output to be a pair of identical uniformly random bits, i.e., \(U = V = \text{Ber}(1/2)\) and introduced what is now known as the Gács-Körner common information of \(P(X, Y)\) to characterize the maximum rate in terms of this quantity. Wyner, on the other hand considered the “inverse” problem where \(X = Y = \text{Ber}(1/2)\) and \(Q\) was arbitrary. Wyner characterized the best
possible conversion rate in this setting in terms of what is now known as the Wyner common information of $Q(U, V)$. There is a rich history of subsequent work (see, for instance, [KA15] and the references within) exploring more general settings where neither $P$ nor $Q$ produces identical copies of some random variable. In such settings, even the question of when can the rate be positive is unknown and this is the question we explore in this chapter.

The Non-Interactive Simulation problem is also a generalization of the Non-Interactive Correlation Distillation problem which was studied by [MO04, MOR⁺06]¹. Our setup can be thought of as a “positive-rate” version of the setup of Gács and Körner. Namely, for a known source distribution $P(X,Y)$, Alice and Bob are given an arbitrary number of i.i.d. samples and wish to generate one sample from the distribution $Q(U, V)$ which is given by $U = V = \text{Ber}(1/2)$. (This is possible if and only if the Gács-Körner rate is positive.)

**Motivation.** Our motivation for studying the best discrete correlation that can be produced is twofold. On the one hand, this question forms part of the landscape of questions arising from a quest to weaken the assumptions about randomness when it is employed in distributed computing. Computational tasks are often solved well if parties have access to a common source of randomness and there has been recent interest in cryptography [AC93, AC98, BS94, CN00, Mau93, RW05], quantum computing [Nie99, CDS08, DB14] and communication complexity [BGI14, CGMS14, GKS16a] to study how the ability to solve these tasks gets affected by weakening the source of randomness. In this space of investigations, it is a very natural question to ask how well one source of randomness can be transformed to a different one, and Non-Interactive Simulation studies exactly this question.

On the other hand, from the analysis point of view, the Non-Interactive Simulation problem forms part of “tensor power” questions that have been challenging to analyze computationally. Specifically, in such questions, the quest is to understand how some quantity behaves as a function of the dimensionality of the problem as the dimension tends to infinity. Notable examples of such problems include the Shannon capacity of a graph [Sha56, Lov79] where the goal is to understand how the independence number of the power of a graph behaves as a function of the exponent. Some more closely related examples arise in the problems of local state transformation of quantum entanglement [Bei12, DB13] and the problem of computing the entangled value of a game (see for e.g., [KKM⁺11] and also the open problem [ope]). A more recent example is the problem of computing the amortized communication complexity of a communication problem. Braverman-Rao [BR11] showed that this equals the information complexity of the communication problem, however the task of approximating the information complexity was only recently shown to be computable [BS15]. In our case, the best non-interactive simulation to get one pair of correlated bits might require many copies of $(x,y)$ drawn from $P$ and the challenge is to determine how many copies get us close. Convergence results of this type are not obvious. Indeed, the task of approximating the Shannon capacity remains open to this day [AL06]. This chapter is motivated in part by the quest to understand tools that can be used to analyze such questions where rate of convergence to the desired quantity is non-trivial to bound.

¹ which considered the problem of maximizing agreement on a single bit, in various multi-party settings.
Estimating Binary Correlations: Previous Work and our Result. In his work generalizing the results of Gács and Körner, Witsenhausen [Wit75] gave an efficient algorithm that achieves a quadratic approximation to the Non-Interactive Simulation problem when \( Q(U,V) \) is the distribution where \( U \) and \( V \) are marginally uniform over \( \pm 1 \) and \( U \) is an \( \rho \)-correlated copy of \( V \), i.e., \( \mathbb{E}[UV] = \rho \) (henceforth, we refer to this distribution as DSBS(\( \rho \))).

Indeed, Witsenhausen introduced the Gaussian correlation problem as an intermediate step to solving this problem and his rounding technique to convert the Gaussian random variables into Boolean ones is essentially the same as that of the Goemans-Williamson algorithm for approximating maximum cut sizes in graphs [GW95]. Already implicit from the work of Witsenhausen is that “maximum correlation” gives a way to upper bound the best achievable \( \rho \) when simulating DSBS(\( \rho \)). Recent works in the information theory community [KA12, KA15, BG15] enhance the collection of analytical tools that can be used to show stronger impossibility results. While these works produce stronger bounds, they do not necessarily converge to the optimal limit and indeed basic questions about simulation remain open. For instance, till our work, even the following question was open [Kam15]: If \( P \) is the uniform distribution on \( \{(0,0),(0,1),(1,0)\} \) and \( Q = \text{DSBS}(0.49) \) (i.e., \( U,V \) are uniformly \( \pm 1 \), with \( \mathbb{E}[UV] = .49 \)), can \( P \) simulate \( Q \) arbitrarily well? This chapter answers such questions in principle. (Specifically, we do give a finite-time procedure to approximate the best \( \rho \) to within arbitrary accuracy. However, we have not run this algorithm to determine the answer to this specific question.)

Below we state our main theorem informally (see Theorem 7.1.3 for the formal statement).

**Theorem 3.1.1 (Informal).** There is an algorithm that takes as inputs a source distribution \( P \), a parameter \( \rho > 0 \) and an error parameter \( \delta > 0 \), runs in time bounded by some computable function of \( P, \rho \) and \( \delta \), and either outputs a non-interactive protocol that simulates DSBS(\( \rho \)) up to additive \( \delta \) in total variation distance, or asserts that there is no protocol that gets \( O(\delta) \)-close to DSBS(\( \rho \)) in total variation distance.

More generally, the proof techniques extend to deciding the non-interactive simulation problem for an arbitrary \( 2 \times 2 \) target distribution. In particular, we also show the following (see Theorem 3.2.3 for the formal statement).

**Theorem 3.1.2 (Informal).** There is an algorithm that takes as inputs a source distribution \( P \), a \( 2 \times 2 \) target distribution \( Q \) and an error parameter \( \delta > 0 \), runs in time bounded by some computable function of \( P, Q \) and \( \delta \), and either outputs a non-interactive protocol that simulates \( Q \) up to additive \( \delta \) in total variation distance, or asserts that there is no protocol that gets \( O(\delta) \)-close to \( Q \) in total variation distance.

The crux of Theorems 3.1.1 and 3.1.2 is to prove computable bounds on the number of copies of \( (X,Y) \) that are needed in order to come \( \delta \)-close to the target distribution. We now describe the challenges towards achieving such bounds, and the techniques we use.

---

Henceforth, we assume that bits are in the set \( \{\pm 1\} \). By a quadratic approximation, we mean an algorithm distinguishing between the cases \( (i) \) \( \rho \geq 1 - \eta \) and \( (ii) \) \( \rho < 1 - O(\sqrt{\eta}) \) for any given parameter \( \eta > 0 \).
**Future Developments**  In a follow-up work, De, Mossel and Neeman showed the decidability of the non-interactive simulation problem for arbitrary *discrete* output alphabets (including the particular case of computing the *noise stability* of a function) [DMN17, DMN18]. In a recent work (that is not included in this thesis) [GKR17], we give a dimension reduction for polynomials over the Gaussian space which we use to significantly improve the bound on the running time of the decidability procedure of [DMN17, DMN18] for non-interactive simulation (and for the particular case of computing the noise stability). Our dimension reduction can be seen as a generalization of the Johnson-Lindenstrauss lemma (which has been extremely influential in computer science with numerous applications including unsupervised learning, compressed sensing, manifold learning, and graph embedding).

### 3.1.1 Proof Overview

We start by describing some illustrative special cases of the problem. In the case where $P = \text{DSBS}(\rho)$, maximal correlation based arguments imply that $\text{DSBS}(\rho)$ is the ‘best’ DSBS distribution that can simulated [Wit75]. Thus, in this case, dictators functions achieve the optimal strategy. Consider now the case where $P$ is a pair of $\rho$-correlated zero-mean unit-variance Gaussians$^3$. Then, Borell’s isoperimetric inequality implies that the strategy where each of Alice and Bob outputs the sign of her/his Gaussian achieves the best possible DSBS [Bor85].

Given the above two examples where a *single-copy* strategy is optimal, it is tempting to try to determine the best DSBS that can be simulated using a single copy of $P$ and hope that it would be close to the optimal DSBS (i.e., to the one that can be simulated using an arbitrary number of copies of $P$). But this approach cannot work as is illustrated by the following example which shows that using many copies of $P$ is in some cases actually needed. Consider the source joint distribution corresponding to the bipartite graph in Figure 3-1 with $\alpha > 0$ being a small parameter (we interpret the distribution as the one obtained by sampling a random edge in the graph). This graph is the union of two components: a low-correlation component which has probability $1 - \alpha$ and a perfect-correlation component which has probability $\alpha$. If we use a small number of copies of $\mu$, the corresponding samples will most likely fall in the low-correlation component, and hence the best DSBS that can be produced in such a way would have a small correlation. On the other hand, as the number of used copies becomes larger than $1/\alpha$, with high probability at least one of the corresponding samples will fall in the perfect-correlation component, and hence the resulting DSBS would have correlation very close to 1. As another example, consider the distribution that is uniform on triples $\{(0, 0), (0, 1), (1, 0)\}$. It follows from [Wit75] that it is possible to simulate $\text{DSBS}(1/3)$ using many copies of this distribution. However, it can be shown that using only a single copy of this distribution (along with private randomness), Alice and Bob can at best simulate $\text{DSBS}(1/4)$.

We now describe at a high level, the main ideas that give us the computable bound on the number of samples of the joint distribution that are sufficient to obtain a $\delta$-approximation to

---

$^3$allowing here continuous distributions for the sake of intuition
a given DSBS(\rho). First, we observe that the problem of deciding if one can come \delta-close to simulating DSBS(\rho), is equivalent to checking if Alice and Bob can non-interactively come up with a distribution \((X,Y)\) on \([-1,1] \times [-1,1]\) such that the marginals of \(X\) and \(Y\) have means close to 0, but \(E[XY]\) is large.

The results on correlation bounds for low-influence functions (obtained using the invariance principle) [MOO05, Mos10], say that if Alice and Bob are using only low-influential functions, then in fact the correlation that they get cannot be much better than that obtained by taking appropriate threshold functions on correlated Gaussians. Moreover, Alice and Bob can in fact simulate correlated Gaussians using only a constant number of samples from the joint distribution, by applying the maximal correlation based technique of Witsenhausen [Wit75].

In the general case, we show that we can first convert Alice and Bob’s functions to have low degree, after which we apply a regularity lemma (inspired from that of [DSTW10]) to conclude that after fixing a constant number of coordinates, the restricted function is in fact low-influential. This reduces the general case to the special case of having low-influential functions and which is handled as described in the previous paragraph.

The more general case of simulating arbitrary \(2 \times 2\) distributions also follows a similar outline. For a more technical overview of the proof, we refer the reader to Section 3.3.1.

### 3.1.2 Organization

In Section 3.2, we give some of the basic definitions. Our main theorems of this chapter are also presented in this section as Theorems 3.2.3 and 7.1.3. In Section 3.3, we state our main technical lemma (Theorem 3.3.1), which is used to prove Theorem 7.1.3. We also give a proof overview for Theorem 3.3.1. In Sections 3.4, 3.5, 3.6 and 3.7, we state and prove the technical lemmas involved in proving Theorem 3.3.1. Finally, in Section 3.8, we put together everything to prove Theorem 3.3.1. We end this chapter by describing some recent follow-up work and discussing some open questions in Section 3.9.
3.2 Preliminaries

3.2.1 Notation

In this chapter, we use script letters $\mathcal{A}$, $\mathcal{B}$, etc. to denote finite sets, and $\mu$ will usually denote a probability distribution. Thus, $(\mathcal{A} \times \mathcal{B}, \mu)$ is a joint probability space. We use $\mu_A$ and $\mu_B$ to denote the marginal distributions of $\mu$. We use letters $x$, $y$, etc. to denote elements of $\mathcal{A}$, and bold letters $\mathbf{x}$, $\mathbf{y}$, etc. to denote elements in $\mathcal{A}^n$. We use $x_i$, $y_i$ to denote individual coordinates of $\mathbf{x}$, $\mathbf{y}$, respectively.

For a probability space $(\mathcal{A}, \mu)$, we will use the following definitions and notations borrowed from [AH11].

- $(\mathcal{A}^n, \mu^\otimes n)$ denotes the product space $\mathcal{A} \times \mathcal{A} \times \cdots \times \mathcal{A}$ endowed with the product distribution.
- $\text{Supp}(\mu) \overset{\text{def}}{=} \{x : \mu(x) > 0\}$ is the support of $\mu$. We would generally assume without loss of generality that $\text{Supp}(\mu) = \mathcal{A}$.
- $\alpha(\mu) \overset{\text{def}}{=} \min \{\mu(x) : x \in \text{Supp}(\mu)\}$ denotes the minimum non-zero probability of any atom in $\mathcal{A}$ under the distribution $\mu$.
- $L^2(\mathcal{A}, \mu)$ denotes the space of functions from $\mathcal{A}$ to $\mathbb{R}$.
- The inner product on $L^2(\mathcal{A}, \mu)$ is denoted by $\langle f, g \rangle_\mu := \mathbb{E}_{x \sim \mu}[f(x)g(x)]$.
- The $\ell_p$-norm is denoted by $\|f\|_p := (\mathbb{E}_{x \sim \mu}|f(x)|^p)^{1/p}$. Also, $\|f\|_\infty := \max_{\mu(x) > 0} |f(x)|$.
- It is easy to verify that $\|f\|_p \leq \|f\|_q$ for any $1 \leq p \leq q$.
- For any two distributions $\mu$ and $\nu$, $d_{TV}(\mu, \nu)$ is the total variation distance between $\mu$ and $\nu$.

3.2.2 The Non-Interactive Simulation Problem

The problem of non-interactive simulation is formally defined as follows.

**Definition 3.2.1 (Non-Interactive Simulation [KA15]).** Let $(\mathcal{A} \times \mathcal{B}, \mu)$ and $(\mathcal{U} \times \mathcal{V}, \nu)$ be two probability spaces. We say that the distribution $\nu$ can be non-interactively simulated using distribution $\mu$, if there exists a sequence of functions $\{f_n\}_{n \in \mathbb{N}}$ and $\{g_n\}_{n \in \mathbb{N}}$ such that $f_n : \mathcal{A}^n \to \mathcal{U}$, $g_n : \mathcal{B}^n \to \mathcal{V}$ and the distribution $\nu_n \sim (f_n(\mathbf{x}), g_n(\mathbf{y}))^{\mu^\otimes n}$ over $\mathcal{U} \times \mathcal{V}$ is such that $\lim_{n \to \infty} d_{TV}(\nu_n, \nu) = 0$.

The notion of non-interactive simulation is pictorially depicted in Figure 3-2. We formulate a natural gap-version of the non-interactive simulation problem defined as follows.
Problem 3.2.2 (Gap-Non-Int-Sim(\(A \times B, \mu\), \((U \times V, \nu), \delta\)). Given probability spaces \((A \times B, \mu)\) and \((U \times V, \nu)\), and an error parameter \(\delta > 0\), distinguish between the following two cases:

(i) There exists a positive integer \(N\), and functions \(f : A^N \to U\) and \(g : B^N \to V\), such that the distribution \(\nu' = (f(x), g(y))_{\mu^N} \) satisfies \(d_{\text{TV}}(\nu', \nu) \leq \delta\).

(ii) For all positive integers \(N\) and all functions \(f : A^N \to U\) and \(g : B^N \to V\), the distribution \(\nu' = (f(x), g(y))_{\mu^N} \) is such that \(d_{\text{TV}}(\nu', \nu) > 8\delta\). \(^4\)

The main result in this chapter is the following theorem showing that the problem of Gap-Non-Int-Sim is decidable when \(|U| = |V| = 2\).

Theorem 3.2.3 (Decidability of Gap-Non-Int-Sim for binary targets). Given probability spaces \((A \times B, \mu)\) and \((U \times V, \nu)\) such that \(|U| = |V| = 2\), and an error parameter \(\delta\), there exists an algorithm that runs in time \(T((A \times B, \mu), \delta)\) (which is an explicitly computable function), and decides the problem of Gap-Non-Int-Sim(\((A \times B, \mu)\), \((U \times V, \nu), \delta\)). The running time \(T((A \times B, \mu), \delta)\) is upper-bounded by

\[
\exp \exp \exp \left( \text{poly} \left( \frac{1}{\delta}, \frac{1}{1 - \rho_0}, \log \left( \frac{1}{\alpha} \right) \right) \right)
\]

\(^4\) for sake of definition, the constant 8 could be replaced by any constant greater than 1. For a minor technical reason however our decidability results (Theorems 3.2.3 and 7.1.3) will require this constant to be strictly greater than 2. We choose to go ahead with 8 for convenience.
where $\rho_0 = \rho(\mathcal{A}, \mathcal{B}; \mu)$ is the maximal correlation of $(\mathcal{A} \times \mathcal{B}, \mu)$ (defined in Section 3.2.6) and $\alpha \overset{\text{def}}{=} \alpha(\mu)$ is the minimum non-zero probability in $\mu$.

**Doubly Symmetric Binary Source**

In order to ease the presentation of ideas in proving the above theorem, we restrict to a special case, where the distribution $(\mathcal{U} \times \mathcal{V}; \nu)$ is a *doubly symmetric binary source* defined below.

**Definition 3.2.4 (Doubly Symmetric Binary Source).** The distribution DSBS($\rho$) is the joint distribution on $\pm 1$ random variables $(U, V)$ given by the following table

<table>
<thead>
<tr>
<th></th>
<th>$V = +1$</th>
<th>$V = -1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$U = +1$</td>
<td>$(1 + \rho)/4$</td>
<td>$(1 - \rho)/4$</td>
</tr>
<tr>
<td>$U = -1$</td>
<td>$(1 - \rho)/4$</td>
<td>$(1 + \rho)/4$</td>
</tr>
</tbody>
</table>

In particular, $\mathbb{E}[U] = \mathbb{E}[V] = 0$ and $\mathbb{E}[UV] = \rho$.

We will prove a special case of Theorem 3.2.3, where the probability space $(\mathcal{U} \times \mathcal{V}, \nu)$ is the distribution DSBS($\rho$) for some $\rho$ (see Theorem 7.1.3 below). Even though we are proving only this special case, the main ideas involved here easily generalize to the proof of Theorem 3.2.3. We give a proof-sketch of this generalization in Section 3.8.1.

**Theorem 3.2.5 (Decidability of Gap-Non-Int-Sim for DSBS targets).** Given a probability space $(\mathcal{A} \times \mathcal{B}, \mu)$, and parameters $\rho$ and $\delta$, there exists an algorithm that runs in time $T((\mathcal{A} \times \mathcal{B}, \mu), \delta)$ (which is an explicitly computable function), and decides the problem of GAP-NON-INT-SIM($(\mathcal{A} \times \mathcal{B}, \mu), \text{DSBS}(\rho), \delta$).

The running time $T((\mathcal{A} \times \mathcal{B}, \mu), \delta)$ is upper-bounded by

\[
\exp \exp \exp \left( \text{poly} \left( \frac{1}{\delta}, \frac{1}{1 - \rho_0}, \log \left( \frac{1}{\alpha} \right) \right) \right)
\]

where $\rho_0 = \rho(\mathcal{A}, \mathcal{B}; \mu)$ is the maximal correlation of $(\mathcal{A} \times \mathcal{B}, \mu)$ (defined in Section 3.2.6) and $\alpha \overset{\text{def}}{=} \alpha(\mu)$ is the minimum non-zero probability in $\mu$.

We will use GAP-NON-INT-SIM($(\mathcal{A} \times \mathcal{B}, \mu), \rho, \delta$) as a shorthand for GAP-NON-INT-SIM($(\mathcal{A} \times \mathcal{B}, \mu), \text{DSBS}(\rho), \delta$). Theorem 7.1.3 will follow easily from the main technical lemma (Theorem 3.3.1). The proof of Theorem 7.1.3, assuming Theorem 3.3.1 is given in Section 3.3.2.

### 3.2.3 Reformulation of GAP-NON-INT-SIM

With the end goal of proving Theorem 7.1.3, we introduce a new problem, Gap-Balanced-Maximum-Inner-Product, to which we show a reduction from GAP-NON-INT-SIM. This new formulation will be better suited for applying our techniques.
Problem 3.2.6 (Gap-Bal-Max-Inner-Product((A × B, μ), ρ, δ)). Given a probability space (A × B, μ), and parameters ρ and δ, distinguish between the following two cases:

(i) There exists a positive integer N, and functions f : A^N → [-1, 1] and g : B^N → [-1, 1] satisfying |E[f(x)]| ≤ δ and |E[g(y)]| ≤ δ, such that the following holds

\[ E[f(x)g(y)] \geq ρ - δ. \]

(ii) For all positive integers N and all functions f : A^N → [-1, 1] and g : B^N → [-1, 1] satisfying |E[f(x)]| ≤ 2δ and |E[g(y)]| ≤ 2δ, the following holds

\[ E[f(x)g(y)] < ρ - 4δ. \]

The following proposition gives a reduction from the Gap-Non-Int-Sim problem to the Gap-Bal-Max-Inner-Product problem.

Proposition 3.2.7. For any probability space (A × B, μ) and ρ, δ > 0, the following reduction holds:

1. Case (i) of Gap-Non-Int-Sim((A × B, μ), ρ, δ) holds \implies Case (i) of Gap-Bal-Max-Inner-Product((A × B, μ), ρ, 2δ) holds.

2. Case (ii) of Gap-Non-Int-Sim((A × B, μ), ρ, δ) holds \implies Case (ii) of Gap-Bal-Max-Inner-Product((A × B, μ), ρ, 2δ) holds.

Proof. Both directions are relatively straightforward.

1. If case (i) of Gap-Non-Int-Sim((A × B, μ), ρ, δ) holds, then there exists a positive integer N and functions f : A^N → {1, -1} and g : B^N → {1, -1} such that the distribution (f(x), g(y))_{μ⊗N} is δ-close to DSBS(ρ) in total variation distance. It follows easily from the definition of total variation distance that |E[f(x)]| ≤ 2δ, |E[g(y)]| ≤ 2δ and E[f(x)g(y)] ≥ ρ - 2δ. These are exactly the conditions needed in case (i) of Gap-Bal-Max-Inner-Product((A × B, μ), ρ, 2δ).

2. We show the contrapositive that if case (ii) of Gap-Bal-Max-Inner-Product((A × B, μ), ρ, 2δ) does not hold, then in fact case (ii) of Gap-Non-Int-Sim((A × B, μ), ρ, δ) also does not hold. Suppose that there exists a positive integer N and functions f : A^N → [-1, 1] and g : B^N → [-1, 1] such that |E[f]| ≤ 4δ, |E[g]| ≤ 4δ and E[f(x)g(y)] ≥ ρ - 8δ. First, we observe that without loss of generality we can assume that E[f(x)g(y)] ≤ ρ. This is because, if that was not the case, we can suitably modify f and g to get f_1 = αf and g_1 = αg such that |E[f_1(x)]| ≤ 4δ, |E[g_1(y)]| ≤ 4δ and E[f_1(x)g_1(y)] = α^2E[f(x)g(y)]. We can choose α suitably such that E[f_1(x)g_1(y)] ≤ ρ.

To show that case (ii) of Gap-Non-Int-Sim((A × B, μ), ρ, δ) does not hold, we obtain randomized functions f' : A^N → {1, -1} and g' : B^N → {1, -1} as follows, f'(x) is equal to 1 with probability (1 + f(x))/2 and -1 otherwise and g'(y) is equal to
1 with probability \((1 + g(y))/2\) and \(-1\) otherwise (The randomness needed can be simulated using some additional copies of \(\mathcal{A}\) and \(\mathcal{B}\)). Note that we have the following: (i) \(\mathbb{E}[f'(x)] = \mathbb{E}[g]\) (ii) \(\mathbb{E}[g'(y)] = \mathbb{E}[g]\) and (iii) \(\rho - 8\delta \leq \mathbb{E}[f'(x)g'(y)] \leq \rho\).

Define \(e_{i,j}\) for \(i, j \in \{1, -1\}\) as follows:

\[
\begin{align*}
  e_{1,1} &= \Pr[f'(x) = +1 \text{ and } g'(y) = +1] - (1 + \rho)/4 \\
  e_{1,-1} &= \Pr[f'(x) = +1 \text{ and } g'(y) = -1] - (1 - \rho)/4 \\
  e_{-1,1} &= \Pr[f'(x) = -1 \text{ and } g'(y) = +1] - (1 - \rho)/4 \\
  e_{-1,-1} &= \Pr[f'(x) = -1 \text{ and } g'(y) = -1] - (1 + \rho)/4
\end{align*}
\]

From (i), (ii) and (iii) above, we have the following:

\[
\begin{align*}
|e_{1,1} + e_{1,-1} - e_{-1,1} - e_{-1,-1}| &\leq 4\delta, \\
|e_{1,1} - e_{1,-1} + e_{-1,1} - e_{-1,-1}| &\leq 4\delta, \\
|e_{1,1} - e_{1,-1} - e_{-1,1} + e_{-1,-1}| &\leq 8\delta.
\end{align*}
\]

In addition, we have that \(e_{1,1} + e_{1,-1} + e_{-1,1} + e_{-1,-1} = 0\). Combining all this, it is easy to infer that \(|e_{i,j}| \leq 4\delta\) for any \(i, j \in \{1, -1\}\). Hence, for \(\nu = (f(x), g(y))_{\mu^\otimes n}\), we get that \(d_{TV}(\nu, \text{DSBS}(\rho)) \leq 8\delta\).

\[\square\]

### 3.2.4 Fourier Analysis and Multi-Linear Polynomials

We recall some background in Fourier analysis that will be useful to us. Let \(q\) be any positive integer and let \((\mathcal{A}, \mu)\) be a finite probability space with \(|\mathcal{A}| = q\). Let \(\mathcal{X}_0, \ldots, \mathcal{X}_{q-1} : \mathcal{A} \to \mathbb{R}\) be an orthonormal basis for the space \(L^2(\mathcal{A}, \mu)\) with respect to the inner product \(\langle \cdot, \cdot \rangle_\mu\). Furthermore, we require that this basis has the property that \(\mathcal{X}_0 = 1\), i.e., the function that is identically 1 on every element of \(\mathcal{A}\).

For \(\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n) \in \mathbb{Z}_q^n\), define \(\mathcal{X}_{\boldsymbol{\sigma}} : \mathcal{A}^n \to \mathbb{R}^n\) as follows,

\[
\mathcal{X}_{\boldsymbol{\sigma}}(x_1, \ldots, x_n) = \prod_{i \in [n]} \mathcal{X}_{\sigma_i}(x_i).
\]

It can be easily seen that the functions \(\{\mathcal{X}_{\boldsymbol{\sigma}}\}_{\boldsymbol{\sigma} \in \mathbb{Z}_q^n}\) form an orthonormal basis for the product space \(L^2(\mathcal{A}^n, \mu^\otimes n)\). Thus, every function \(f \in L^2(\mathcal{A}^n, \mu^\otimes n)\) can be written as

\[
f(x) = \sum_{\boldsymbol{\sigma} \in \mathbb{Z}_q^n} \hat{f}(\boldsymbol{\sigma}) \mathcal{X}_{\boldsymbol{\sigma}}(x)
\]

where \(\hat{f} : \mathbb{Z}_q^n \to \mathbb{R}\) can be obtained as \(\hat{f}(\boldsymbol{\sigma}) = \langle f, \mathcal{X}_{\boldsymbol{\sigma}} \rangle_\mu\). The function \(\hat{f}\) is the Fourier transform of \(f\) with respect to the basis \(\{\mathcal{X}_{\sigma}\}_{\sigma \in \mathbb{Z}_q}\). Although we will work with an arbitrary
(albeit fixed) basis, many of the important properties of the Fourier transform are basis-independent. The most basic properties of $\hat{f}$ are summarized in the following fact which follows from the orthonormality of $\{X_\sigma\}_{\sigma \in \mathbb{Z}_q^n}$.

**Fact 3.2.8.** We have:

- **Plancherel’s Identity**: $E[f(x)g(x)] = \sum_\sigma \hat{f}(\sigma)\hat{g}(\sigma)$.
- **As a special case, we have Parseval’s identity**, $E[f(x)^2] = \sum_\sigma \hat{f}(\sigma)^2$.
- $E[f] = \hat{f}(0)$.
- $\text{Var}[f] = \sum_{\sigma \neq 0} \hat{f}(\sigma)^2$.

In this chapter, we will deal with joint probability spaces of the type $(\mathcal{A} \times \mathcal{B}, \mu)$. In such cases, we will denote the marginal probability spaces as $(\mathcal{A}, \mu_A)$ and $(\mathcal{B}, \mu_B)$. We will abuse notations to use $X_\sigma$ to denote the orthonormal basis vectors for both $L_2(\mathcal{A}_n, \mu_\otimes^n_A)$ as well as $L_2(\mathcal{B}_n, \mu_\otimes^n_B)$. The space of $\sigma$ will be $\mathbb{Z}_{|\mathcal{A}|}^n$ or $\mathbb{Z}_{|\mathcal{B}|}^n$ accordingly, and will be clear from context.

For $\sigma \in \mathbb{Z}_q^n$, the degree of $\sigma$ is denoted by $|\sigma| \overset{\text{def}}{=} |\{i \in [n] : \sigma_i \neq 0\}|$. We say that the degree of a function $f \in L_2(\mathcal{A}_n, \mu_\otimes^n)$, denoted by $\text{deg}(f)$, is the largest value of $|\sigma|$ such that $\hat{f}(\sigma) \neq 0$.

**Definition 3.2.9 (Influence).** For every coordinate $i \in [n]$, $\text{Inf}_i(f)$ is the $i$-th influence of $f$, and $\text{Inf}(f)$ is the total influence, which are defined as

$$\text{Inf}_i(f) \overset{\text{def}}{=} E_{x \rightarrow i} \left[ \text{Var}_i[f(x)] \right] \quad \text{Inf}(f) \overset{\text{def}}{=} \sum_{i=1}^n \text{Inf}_i(f).$$

The basic properties of influence are summarized in the following fact.

**Fact 3.2.10.** For any function $f \in L_2(\mathcal{A}_n, \mu_\otimes^n)$, we have the following:

(i) $\text{Inf}_i(f) = \sum_{\sigma, \sigma_i \neq 0} \hat{f}(\sigma)^2$ and hence, for all $i$, $\text{Inf}_i(f) \leq \text{Var}(f)$.

(ii) $\text{Inf}(f) = \sum |\sigma| \cdot \hat{f}(\sigma)^2$.

(iii) If $\text{deg}(f) = d$, then $\text{Inf}(f) \leq d \cdot \text{Var}[f]$.

---

5 We will interchangeably use the word *polynomial* to talk about any function in $L_2(\mathcal{A}_n, \mu_\otimes^n)$. 

71
Restrictions of Polynomials

We will often use restrictions of polynomials. For any subset \( H \subseteq [n] \), we will use \( x_H \) to denote the tuple of variables in \( x \) with indices in \( H \). For any function \( P \in L^2(\mathcal{A}^n, \mu^\otimes n) \), and any \( \xi \in \mathcal{A}^H \), we will use \( P_\xi \) to denote the function obtained by restriction of \( x_H \) to \( \xi \), that is, \( P_\xi(x_T) = P(x_H \leftarrow \xi, x_T) \) (where \( T = [n] \setminus H \)); whenever we use such terminology, the subset \( H \) will be clear from context. We will use the phrase “\( \xi \) fixes \( H \) over \( \mathcal{A} \)” to mean such a restriction. We will use \( \{\sigma_H\} \) to denote all degree sequences in \( \mathbb{Z}_q^H \), and similarly \( \{\sigma_T\} \) to denote all degree sequences in \( \mathbb{Z}_q^T \). We use \( \sigma_H \circ \sigma_T \) to denote \( \sigma \in \mathbb{Z}_q^n \) such that \( \sigma_i = (\sigma_H)_i \) if \( i \in H \) or \( (\sigma_T)_i \) if \( i \in T \).

We now state a lemma that will be needed,

**Lemma 3.2.11** (cf. Lemma 3.3 in [DSTW10]). For any function \( P \in L^2(\mathcal{A}^n, \mu^\otimes n) \), consider a random assignment \( \xi \sim \mu^H \) to the variables \( x_H \). Let \( T = [n] \setminus H \). Then, for all \( i \in T \), it holds that \( \mathbb{E}_{\xi}[\text{Inf}_i(P_\xi)] = \text{Inf}_i(P) \). Also, \( \mathbb{E}_{\xi}[\text{Var}(P_\xi)] \leq \text{Var}(P) \).

To prove the lemma, we first recall the following fact about the expected value of Fourier coefficients under random restrictions.

**Fact 3.2.12.** Let \( P \in L^2(\mathcal{A}^n, \mu^\otimes n) \). For any subset \( H \subseteq [n] \), consider an assignment \( \xi \) to the variables \( x_H \). Let \( T = [n] \setminus H \). Then, we have that

\[
\hat{P}_\xi(\sigma_T) = \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T) \cdot \mathcal{X}_{\sigma_H}(\xi)
\]

and therefore

\[
\mathbb{E}_{\xi} \left[ \hat{P}_\xi(\sigma_T)^2 \right] = \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T)^2.
\]

**Proof.** The first part follows from simply substituting \( P_\xi(x_T) = P(x_H \leftarrow \xi, x_T) \), and taking the inner product with \( \mathcal{X}_{\sigma_T}(x_T) \):

\[
\hat{P}_\xi(\sigma_T) = \left\langle \sum_{\sigma_H \circ \sigma_T'} \hat{P}(\sigma_H \circ \sigma_T') \cdot \mathcal{X}_{\sigma_H}(\xi)\mathcal{X}_{\sigma_T'}(x_T), \mathcal{X}_{\sigma_T}(x_T) \right\rangle = \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T) \cdot \mathcal{X}_{\sigma_H}(\xi).
\]

The second part simply follows from the orthonormality of the characters \( \mathcal{X}_{\sigma_H} \) and \( \mathcal{X}_{\sigma_T} \).
for $\sigma_H \neq \sigma'_H$. In particular, we have the following:

$$
\mathbb{E}_\xi \left[ \hat{P}_\xi(\sigma_T)^2 \right] = \mathbb{E}_\xi \left[ \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi)^2 \right] = \mathbb{E}_\xi \left[ \sum_{\sigma_H, \sigma'_H} \hat{P}(\sigma_H \circ \sigma_T) \cdot \hat{P}(\sigma'_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) \cdot \chi_{\sigma'_H}(\xi) \right] = \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T)^2.
$$

Intuitively, the above facts say that all the Fourier weight on degree sequences $\{\sigma_H \circ \sigma_T\}_{\sigma_H}$ collapses down onto $\sigma_T$ in expectation. Consequently, the influence of an unrestricted variable does not change, and the variance does not increase in expectation under random restrictions, as both these quantities are sums of Fourier weight on certain $\sigma_T$’s.

Proof of Lemma 3.2.11. We simply use Facts 3.2.8 and 3.2.10 in addition to Fact 3.2.12 to prove the lemma. Namely, from Facts 3.2.8 and 3.2.12, we get that

$$
\mathbb{E}_\xi[\text{Var}(P_\xi)] = \mathbb{E}_\xi \left[ \sum_{\sigma_T \neq 0} \hat{P}_\xi(\sigma_T)^2 \right] = \sum_{\sigma_T \neq 0} \mathbb{E}_\xi \left[ \hat{P}_\xi(\sigma_T)^2 \right] = \sum_{\sigma_T \neq 0} \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T)^2 \leq \text{Var}(P).
$$

Similarly, from Facts 3.2.10 and 3.2.12, we get that for all $i \in T$,

$$
\mathbb{E}_\xi[\text{Inf}_i(P_\xi)] = \mathbb{E}_\xi \left[ \sum_{\sigma_T \neq 0} \left( \prod_{(\sigma_T)_i = 0} \hat{P}_\xi(\sigma_T) \right)^2 \right] = \sum_{\sigma_T \neq 0} \mathbb{E}_\xi \left[ \hat{P}_\xi(\sigma_T)^2 \right] = \sum_{\sigma_T \neq 0} \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T)^2 = \text{Inf}_i(P).
$$

### 3.2.5 Hypercontractivity and Moment Bounds

The following moment bound for low-degree polynomials appears as Theorem 2.7 in [AH11], which in turn follows from hypercontractivity.

**Theorem 3.2.13 ([Wol07]).** Let $(\mathcal{A}, \mu)$ be a finite probability space in which the minimum non-zero probability is $\alpha(\mu) \leq \frac{1}{2}$. Then, for $p \geq 2$, every degree-$d$ polynomial $f \in L^2(\mathcal{A}^n, \mu^\otimes n)$
satisfies
\[ \|f\|_p \leq C_p(\alpha)^{d/2} \|f\|_2. \]
Here, \( C_p \) is defined by
\[ C_p(\alpha) = \frac{A^{1/p'} - A^{-1/p'}}{A^{1/p} - A^{-1/p}} \]
where \( A = (1-\alpha)/\alpha \) and \( 1/p + 1/p' = 1 \). The value at \( \alpha = 1/2 \) is taken to be the limit of the above expression as \( \alpha \to 1/2 \), i.e., \( C_p(1/2) = p - 1 \).

We will use the following known concentration bound for low-degree polynomials.

**Theorem 3.2.14** ([AH11]; Theorem 2.12). Let \( f \in L^2(\mathcal{A}^n, \mu^\otimes n) \) be a degree-\( d \) polynomial. Then, for any \( t > e^{d/2} \),
\[ \Pr[|f| > t \cdot \|f\|_2] \leq \exp(-ct^{2/d}) \]
where \( c := \frac{\alpha(p)d}{e} \).

**Definition 3.2.15** (Bonami-Beckner Operator). For any \( \rho \in [0, 1] \), the Bonami-Beckner operator \( T_\rho \) on a probability space \( (\mathcal{A}, \mu) \) is given by its action on any \( f : \mathcal{A} \to \mathbb{R} \), as follows:
\[ (T_\rho f)(x) = \mathbb{E}[f(Y)|X = x] \]
where the conditional distribution of \( Y \) given \( X = x \) is \( \rho \delta_x + (1 - \rho)\mu \) where \( \delta_x \) is the delta measure on \( x \). In other words, given \( X = x \), \( Y \) is obtained by either setting it to \( x \) with probability \( \rho \) or independently sampling from \( \mu \) with probability \( (1 - \rho) \).

For the product space \( (\mathcal{A}^n, \mu^\otimes n) \), we define the Bonami-Beckner operator \( T_\rho \) as \( T_\rho = \otimes_{i=1}^n T_\rho^{(i)} \), where \( T_\rho^{(i)} \) is the Bonami-Beckner operator on the \( i \)-th coordinate \( (\mathcal{A}, \mu) \).

### 3.2.6 Maximal Correlation and Witsenhausen’s Rounding

The “maximal correlation coefficient” was first introduced by Hirschfeld [Hir35] and Gebelein [Geb41] and then studied by Rényi [Rén59].

**Definition 3.2.16** (Maximal Correlation). Given a joint probability space \( (\mathcal{A} \times \mathcal{B}, \mu) \), we define the maximal correlation of the joint distribution \( \rho(\mathcal{A}, \mathcal{B}; \mu) \) as follows:
\[ \rho(\mathcal{A}, \mathcal{B}; \mu) = \sup \left\{ \mathbb{E}_{(x,y) \sim \mu}[f(x)g(y)] \middle| f : \mathcal{A} \to \mathbb{R}, \quad g : \mathcal{B} \to \mathbb{R}, \quad \mathbb{E}[f] = \mathbb{E}[g] = 0, \quad \text{Var}(f) = \text{Var}(g) = 1 \right\}. \]
Maximal correlation has the following properties which imply necessary conditions for when non-interactive simulation is possible!

**Fact 3.2.17** (Properties of Maximal Correlation (cf. [BDK05])).

1. (Tensorization) : For all joint probability spaces \( (\mathcal{A} \times \mathcal{B}, \mu) \), it is the case that \( \rho(\mathcal{A}^n, \mathcal{B}^n; \mu^\otimes n) = \rho(\mathcal{A}, \mathcal{B}; \mu) \).
2. (Data Processing) : For all joint probability spaces \((A \times B, \mu)\), and any functions \(f : A \to U\) and \(g : B \to V\), it is the case that \(\rho(A, B; \mu) \geq \rho(U, V; \nu)\), where \(\nu\) is the distribution \((f(x), g(y))_{(x,y) \sim \mu}\).

3. (Lower Semi-Continuous) : If distributions \((U \times V; \nu_n)\) are such that \(\lim_{n \to \infty} \nu_n = \nu\), then \(\lim_{n \to \infty} \rho(U, V; \nu_n) \geq \rho(U, V; \nu)\).

**Corollary 3.2.18** (Necessary Condition for Non-Interactive Simulation). Let \((A \times B, \mu)\) and \((U \times V, \nu)\) be two probability spaces. If the distribution \(\nu\) can be non-interactively simulated using distribution \(\mu\), then \(\rho(A, B; \mu) \geq \rho(U, V; \nu)\).

A simple fact that can be easily verified is that the maximal correlation of the distribution \(DSBS(\rho)\) is \(\rho\). And hence if \((A \times B, \mu)\) can non-interactively simulate \(DSBS(\rho^*)\), then \(\rho^* \leq \rho(A, B; \mu)\). In addition, Witsenhausen [Wit75] showed that any joint probability space \((A \times B, \mu)\) can simulate \(DSBS(\rho^*)\) for \(\rho^* = 1 - \frac{2\arccos(\rho(A, B; \mu))}{\pi}\). All together, we have the following theorem,

**Theorem 3.2.19** (Witsenhausen [Wit75]). For any joint probability space \((A \times B, \mu)\), with maximal correlation \(\rho = \rho(A, B; \mu)\), the largest \(\rho^*\) for which \((A \times B, \mu)\) can non-interactively simulate \(DSBS(\rho^*)\) is bounded as follows:

\[
1 - \frac{2\arccos(\rho)}{\pi} \leq \rho^* \leq \rho.
\]

Note that maximal correlation is an efficiently computable quantity, namely, it is the second largest singular value of the Markov operator\(^6\) corresponding to \((A \times B, \mu)\).

**Remark 3.2.20.** The astute reader might have noticed a strong resemblance between Theorem 3.2.19 and the random hyperplane rounding of Goemans-Williamson [GW95] used in the approximation algorithm for MAX-CUT. This is not a coincidence and indeed the bounds in Theorem 3.2.19 come from morally the same technique as in [GW95].

In this context, we will use the following shorthand for a \(\rho\)-correlated two-dimensional Gaussian.

**Definition 3.2.21** (Two-Dimensional Gaussian). Let \(G(\rho)\) denote a two-dimensional Gaussian distribution with mean \(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\) and covariance matrix \(\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\).

### 3.2.7 Two-Dimensional Berry-Esseen Theorem

We will need the following two-dimensional Berry-Esseen theorem. The proof is very similar to Theorem 68 of [MORS10]. The main difference is that in our case the random variables are not necessarily binary-valued, but they do have finite support. We include the proof for completeness.

---

\(^6\)The Markov operator corresponding to \((A \times B, \mu)\) is a \(|A| \times |B|\) matrix \(T\) which is given by \(T(x, y) = \mu(y | X = x)\).
Lemma 3.2.22 (Two-Dimensional Berry-Esseen). Let \((X, Y)\) be any pair of correlated real-valued random variables with finite support such that, \(\mathbb{E}[X] = \mathbb{E}[Y] = 0\) and \(\text{Var}(X) = \text{Var}(Y) = 1\) and \(\mathbb{E}[XY] = \rho\). For every \(\zeta > 0\), there exists \(w \triangleq w((X, Y), \zeta) \in \mathbb{N}\), such that for every \(a, b \in \mathbb{R}\), it is the case that,

\[
| \Pr[X \leq a, Y \leq b] - \Pr[G_1 \leq a, G_2 \leq b] | \leq \zeta
\]

where \(X = \sum_{i=1}^{w} X_i / \sqrt{w}\), \(Y = \sum_{i=1}^{w} Y_i / \sqrt{w}\) (with \((X_i, Y_i)\) drawn i.i.d. from \((X, Y)\)) and \((G_1, G_2) \sim \mathcal{G}(\rho)\).

In particular, one may take \(w = O\left(\frac{1 + \rho}{\alpha(1-\rho)^3 \zeta^2}\right)\), where \(\alpha\) is the minimum non-zero probability in the distribution \((X, Y)\).

In order to prove Lemma 3.2.22, we need the following statement that appears as Theorem 16 in [KKMO07] and as Corollary 16.3 in [BR86].

Theorem 3.2.23. Let \(Z_1, \ldots, Z_w\) be independent random variables taking values in \(\mathbb{R}^k\) and satisfying:

- \(\mathbb{E}[Z_j]\) is the all-zero vector for every \(j \in \{1, \ldots, w\}\).
- \(\sum_{j=1}^{w} \text{Cov}[Z_j] / w = V\) where \(\text{Cov}\) denotes the covariance matrix.
- \(\lambda\) is the smallest eigenvalue of \(V\) and \(\Lambda\) is the largest eigenvalue of \(V\).
- \(\sum_{j=1}^{w} \mathbb{E} \left[ \|Z_j\|^3 \right] / w = \rho_3 < \infty\).

Let \(Q_w\) denote the distribution of \((Z_1 + \cdots + Z_w) / \sqrt{w}\), let \(\Phi_{0,V}\) denote the distribution of the \(k\)-dimensional Gaussian with mean 0 and covariance matrix \(V\), and let \(\eta = C \lambda^{-3/2} \rho_3 w^{-1/2}\), where \(C\) is a certain universal constant. Then, for any Borel set \(A\),

\[
| Q_w(A) - \Phi_{0,V}(A) | \leq \eta + B(A)
\]

where \(B(A)\) is the following measure of the boundary of \(A\): \(B(A) = 2 \sup_{y \in \mathbb{R}^k} \Phi_{0,V}((\partial A)^\eta' + y)\), \(\eta' = \Lambda^{1/2} \eta\) and \((\partial A)^\eta'\) denotes the set of points within distance \(\eta'\) of the topological boundary of \(A\).

Proof of Lemma 3.2.22. We apply Theorem 5.9.2 with \(k = 2\). Let \(Z = (X, Y)\), and hence we have that,

\[
\mathbb{E}[Z] = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \text{Cov}[Z] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.
\]

Let \(Z_i = (X_i, Y_i)\). Since all \(Z_i\) are i.i.d. distributed according to \(Z\), we have that \(V = \sum_{j=1}^{w} \text{Cov}[(X_j, Y_j)] / w\) is also \(\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\). It follows that the smallest and largest eigenvalues of
$V$ are $\lambda = 1 - \rho$ and $\Lambda = 1 + \rho$ respectively. Moreover, since the underlying distribution has finite support, we have that

$$\sum_{j=1}^{w} \frac{\mathbb{E}[\|Z_j\|^3]}{w} = \mathbb{E}[\|Z\|^3] < \max Z \cdot \mathbb{E}[\|Z\|^2] \leq \frac{1}{\sqrt{\alpha}},$$

(where $\alpha$ is the smallest atom in the distribution $(X, Y)$). Thus, we get that $\rho_3 \leq 1/\sqrt{\alpha}$. Hence, $\eta = O((1 - \rho)^{-3/2}\alpha^{-1/2}w^{-1/2})$. As in [KKMO07], one can check that the topological boundary of any set of the form $(-\infty, a] \times (-\infty, b]$ is $O(\eta')$, where $\eta' = (1 + \rho)^{1/2}$. Thus, from Lemma 3.2.22 it follows by choosing $w$ sufficiently large so that $O((1 + (1 + \rho)^{1/2})(1 - \rho)^{-3/2}\alpha^{-1/2}w^{-1/2}) \leq \zeta$.

In particular, it suffices to choose $w = O\left(\frac{1 + \rho}{\alpha(1 - \rho)^3\xi^2} \right)$. As in [KKMO07], one can check that the topological boundary of any set of the form $(a, \infty) \times (b, \infty)$ is $O(\zeta')$, where $\zeta' = (1 + \rho)^{1/2}$.

Thus, from Lemma 3.2.22 it follows by choosing $w$ sufficiently large so that $O\left(\frac{1 + \rho}{\alpha(1 - \rho)^3\xi^2} \right) \leq \zeta$.

In particular, it suffices to choose $w = O\left(\frac{1 + \rho}{\alpha(1 - \rho)^3\xi^2} \right)$. As in [KKMO07], one can check that the topological boundary of any set of the form $(a, \infty) \times (b, \infty)$ is $O(\zeta')$, where $\zeta' = (1 + \rho)^{1/2}$.

Thus, from Lemma 3.2.22 it follows by choosing $w$ sufficiently large so that $O\left(\frac{1 + \rho}{\alpha(1 - \rho)^3\xi^2} \right) \leq \zeta$.

### 3.3 Main Technical Lemma and Overview

In this section, we state the main technical lemma which will be used to solve the Gap-Bal-Max-Inner-Product problem. We also give a high-level overview of the proof techniques.

**Theorem 3.3.1.** Given any joint probability space $(\mathcal{A} \times \mathcal{B}, \mu)$ and any $\delta > 0$, there exists a positive integer $n_0 = n_0((\mathcal{A} \times \mathcal{B}, \mu), \delta)$ such that for any positive integer $n$ and any functions $f : \mathcal{A}^n \to [-1, 1]$ and $g : \mathcal{B}^n \to [-1, 1]$, there exist functions $\tilde{f} : \mathcal{A}^{n_0} \to [-1, 1]$ and $\tilde{g} : \mathcal{B}^{n_0} \to [-1, 1]$ such that $|\mathbb{E}[\tilde{f}] - \mathbb{E}[f]| \leq \delta/3$, $|\mathbb{E}[\tilde{g}] - \mathbb{E}[g]| \leq \delta/3$ and

$$\mathbb{E}_{(x, y) \sim \mu^{\otimes n}} [\tilde{f}(x) \cdot \tilde{g}(y)] \geq \mathbb{E}_{(x, y) \sim \mu^{\otimes n}} [f(x) \cdot g(y)] - \delta.$$

Most importantly, $n_0$ is a computable function of the parameters of the problem. In particular, one may take,

$$n_0 = \exp \left( \text{poly} \left( \frac{1}{\delta}, \frac{1}{1 - \rho_0}, \log \left( \frac{1}{\alpha} \right) \right) \right)$$

where $\rho \overset{\text{def}}{=} \rho(\mathcal{A}, \mathcal{B}; \mu)$ is the maximal correlation of $(\mathcal{A} \times \mathcal{B}, \mu)$ and $\alpha \overset{\text{def}}{=} \alpha(\mu)$ is the minimum non-zero probability in $\mu$.

#### 3.3.1 Proof Overview

The proof of Theorem 3.3.1 goes through a series of intermediate steps, which we describe at a high-level here. At each step, we lose only a small amount in the correlation. The first three steps preserve the marginals $\mathbb{E}[f]$ and $\mathbb{E}[g]$ exactly, while the fourth step incurs a small additive error in the same. The full proof is presented in Section 3.8.
Step 1: **Smoothing of Strategies.** We transform $f$ and $g$ into functions $f_1$, $g_1$ such that $f_1$ and $g_1$ have ‘most’ of their Fourier mass concentrated on terms of degree at most $d$, where $d$ is a constant that depends on the distribution $(\mathcal{A} \times \mathcal{B}, \mu)$ and a tolerance parameter, but is independent of $n$. This transformation is described in Section 3.4.

Step 2: **Regularity Lemma for Low-Degree Functions.** We first prove a regularity lemma (similar to the one in [DSTW10]) which roughly shows that for any degree-$d$ polynomial, there exists an $h$-sized subset of variables, such that under a random restriction of the variables in this subset, the resulting function on the remaining variables has low individual influences (i.e. $\leq \tau$). Note that $h$ will be a constant depending on the degree $d$ and $\tau$, but will be independent of $n$.

We apply this regularity lemma on the degree-$d$ truncated versions of both $f_1$ and $g_1$ obtained from Step 1. We take the union of the subsets obtained for $f_1$ and $g_1$. We show that with high probability over random restrictions of the variables in this subset, the resulting restriction of $f_1$ and $g_1$ on the remaining variables has low individual influences. This step is described in Section 3.5.

Note that this step does not change the functions $f_1$ and $g_1$ at all, but we gain some structural knowledge about the same.

Step 3: **Correlation Bounds for Low-Influence Functions.** We use results about correlation bounds for low influential functions [MOO05, Mos10]. Intuitively, these results suggest that if the functions $f_1$ and $g_1$ were low-influential functions to begin with, then the correlation $\mathbb{E}[f_1(x)g_1(y)]$ will not be ‘much’ better than the correlation between certain threshold functions applied on correlated Gaussians.

We apply the above correlation bounds for the low-influential functions obtained by restrictions of the small subset of variables in $f_1$ and $g_1$, to obtain functions $f_2 : \mathcal{A}^h \times \mathbb{R} \rightarrow [-1, 1]$ and $g_2 : \mathcal{B}^h \times \mathbb{R} \rightarrow [-1, 1]$, where Alice and Bob together have access to $h$ samples from $(\mathcal{A} \times \mathcal{B}, \mu)$ and a single copy of $\rho$-correlated Gaussians, that is, $\mathcal{G}(\rho)$ (see Definition 3.2.21). Here, the correlation $\rho$ is the same as the maximal correlation $\rho(\mathcal{A}, \mathcal{B}; \mu)$. This step is described in Section 3.6.

Step 4: **Simulating Correlated Gaussians.** Finally, Alice and Bob can non-interactively simulate the distribution $\mathcal{G}(\rho)$ using constantly many samples from $(\mathcal{A} \times \mathcal{B}, \mu)$. This is done using the technique of Witsenhausen [Wit75], which primarily uses a two-dimensional central limit theorem. This step is described in Section 3.7.

### 3.3.2 Decidability of $\text{Gap-Non-Int-Sim}$

Assuming Theorem 3.3.1, we now give the algorithm as described in Theorem 7.1.3.

**Proof of Theorem 7.1.3.** We have from Proposition 3.2.7 that in order to decide the $\text{Gap-Non-Int-Sim}((\mathcal{A} \times \mathcal{B}, \mu), \rho, \delta)$ problem, it suffices to decide the $\text{Gap-Bal-Max-Inner-Product}((\mathcal{A} \times \mathcal{B}, \mu), \rho, 2\delta)$ problem.
If we were in the YES case of $\text{Gap-Bal-Max-Inner-Product}((\mathcal{A} \times \mathcal{B}, \mu), \rho, 2\delta)$, then there exists a positive integer $n$ and functions $f : \mathcal{A}^n \to [-1, 1]$ and $g : \mathcal{B}^n \to [-1, 1]$, such that $|\mathbb{E}[f(x)]| \leq 2\delta$, $|\mathbb{E}[g(y)]| \leq 2\delta$ and $\mathbb{E}[f(x) \cdot g(y)] \geq \rho - 2\delta$. Using Theorem 3.3.1 with parameter $\delta$, we get that there exist functions $\tilde{f} : \mathcal{A}^n \to [-1, 1]$ and $\tilde{g} : \mathcal{B}^n \to [-1, 1]$ such that $|\mathbb{E}[\tilde{f}(x)]| \leq 8\delta/3$, $|\mathbb{E}[\tilde{g}(y)]| \leq 8\delta/3$ and $\mathbb{E}[\tilde{f}(x) \cdot \tilde{g}(y)] \geq \rho - 3\delta$.

In the NO case of $\text{Gap-Bal-Max-Inner-Product}((\mathcal{A} \times \mathcal{B}, \mu), \rho, 2\delta)$, we have that for all positive integers $n$, in particular for $n = n_0$, there do not exist functions $f : \mathcal{A}^n \to [-1, 1]$ and $g : \mathcal{B}^n \to [-1, 1]$ such that $|\mathbb{E}[f(x)]| \leq 4\delta$, $|\mathbb{E}[g(y)]| \leq 4\delta$ and $\mathbb{E}[f(x) \cdot g(y)] \geq \rho - 8\delta$.

This naturally gives us a brute force algorithm: Analyze all possible functions $\tilde{f} : \mathcal{A}^{n_0} \to [-1, 1]$ and $\tilde{g} : \mathcal{B}^{n_0} \to [-1, 1]$ to check if there exist functions satisfying $|\mathbb{E}[\tilde{f}(x)]| \leq 8\delta/3$, $|\mathbb{E}[\tilde{g}(y)]| \leq 8\delta/3$ and $\mathbb{E}[\tilde{f}(x) \cdot \tilde{g}(y)] \geq \rho - 3\delta$. For purposes of our algorithm we can treat the range $[-1, 1]$ as a discrete set $R \stackrel{\text{def}}{=} \{k \delta^2 / 10 : k \in \mathbb{Z}, |k| < 10/\delta^2\}$. This ensures that if indeed such a desired $\tilde{f}$ and $\tilde{g}$ exist, then we will find functions $\tilde{f}' : \mathcal{A}^{n_0} \to R$ and $\tilde{g}' : \mathcal{B}^{n_0} \to R$ such that $|\mathbb{E}[\tilde{f}'(x)]|, |\mathbb{E}[\tilde{g}'(y)]| \leq 8\delta/3 + O(\delta^2)$ and $\mathbb{E}[\tilde{f}'(x) \cdot \tilde{g}'(y)] \geq \rho - 3\delta - O(\delta^2)$. In the YES case, we will find such functions, whereas in the NO case, $\tilde{f}'$ and $\tilde{g}'$ as above simply do not exist.

It is easy to see that this brute force can be done in $(1/\delta)^{O((|\mathcal{A}|+|\mathcal{B}|)^{n_0})}$ time, which is upper bounded by the running time claimed in Theorem 7.1.3.

### 3.4 Smoothing of Strategies

The first step in our approach is to obtain smoothed versions of the functions $f : \mathcal{A}^n \to [-1, 1]$ and $g : \mathcal{B}^n \to [-1, 1]$, which have small Fourier tails, without hurting the correlation by much. In particular, we show the following lemma.

**Lemma 3.4.1 (Smoothing of Strategies).** Given any joint probability space $(\mathcal{A} \times \mathcal{B}, \mu)$ and parameters $\lambda, \eta > 0$, there exists a positive integer $d = d((\mathcal{A} \times \mathcal{B}, \mu), \lambda, \eta)$ such that for any positive integer $n$ and any functions $f : \mathcal{A}^n \to [-1, 1]$ and $g : \mathcal{B}^n \to [-1, 1]$, there exist functions $f_1 : \mathcal{A}^n \to [-1, 1]$ and $g_1 : \mathcal{B}^n \to [-1, 1]$ such that $\mathbb{E}[f_1] = \mathbb{E}[f]$, $\mathbb{E}[g_1] = \mathbb{E}[g]$, and

$$
|\mathbb{E}_{(x,y) \sim \mu^\otimes n}[f_1(x) \cdot g_1(y)] - \mathbb{E}_{(x,y) \sim \mu^\otimes n}[f(x) \cdot g(y)]| \leq \lambda
$$

where $f_1$ and $g_1$ have low energy Fourier tails, i.e.,

$$
\sum_{|\sigma| > d} \widehat{f_1}(\sigma)^2 \leq \eta \quad \text{and} \quad \sum_{|\sigma| > d} \widehat{g_1}(\sigma)^2 \leq \eta.
$$

In particular, one may take $d = \frac{\log \eta}{2 \log \gamma}$, where $\gamma = 1 - C \frac{(1-\rho)^2}{\log(1/\lambda)}$, and $\rho = \rho(\mathcal{A}, \mathcal{B}; \mu)$.

To prove Lemma 3.4.1, we use Lemma 6.1 of Mossel [Mos10]. We state a specialized version of Mossel’s lemma, which suffices for our application.
Lemma 3.4.2 ([Mos10]). Let \((\mathcal{A} \times \mathcal{B}, \mu)\) be a finite joint probability space, such that \(\rho(\mathcal{A} \times \mathcal{B}, \mu) \leq \rho\).

Let \(P \in L^2(\mathcal{A}^n, \mu_{A}^{\otimes n})\) and \(Q \in L^2(\mathcal{B}^n, \mu_{B}^{\otimes n})\) be multi-linear polynomials. Let \(\epsilon > 0\) and \(\gamma\) be chosen sufficiently close to 1 so that,

\[\gamma \geq (1 - \epsilon)^{\log \rho / (\log \epsilon + \log \rho)}\.

Then,

\[|\mathbb{E}[P(x)Q(y)] - \mathbb{E}[T_{\gamma}P(x)T_{\gamma}Q(y)]| \leq 2\epsilon \operatorname{Var}[P] \operatorname{Var}[Q].\]

In particular, there exists an absolute constant \(C\) such that it suffices to take

\[\gamma \overset{\text{def}}{=} 1 - C(1 - \rho)\epsilon / \log(1 / \epsilon).

Proof of Lemma 3.4.1. Given parameters \(\lambda\) and \(\eta\), we first choose \(\epsilon\) and \(\gamma\) in Lemma 3.4.2, such that \(\epsilon = \lambda / 2\) and \(\gamma = 1 - C ((1 - \rho)\epsilon / (\log(1 / \epsilon))\) as required. Then, we choose \(d\) to be large enough so that \(\gamma^{2d} \leq \eta\), that is, \(d = (\log \eta) / (2 \log \gamma)\). Now, given functions \(f : \mathcal{A}^n \to [-1, 1]\) and \(g : \mathcal{B}^n \to [-1, 1]\), we obtain functions \(f_1\) and \(g_1\) as follows: \(f_1(x) = T_{\gamma}f(x)\) and \(g_1(y) = T_{\gamma}g(y)\). It is easy to see that, \(\mathbb{E}[f_1(x)] = \mathbb{E}[f(x)]\) and \(\mathbb{E}[g_1(y)] = \mathbb{E}[g(y)]\). From Lemma 3.4.2, and the fact that \(\operatorname{Var}[f], \operatorname{Var}[g] \leq 1\), we get \(|\mathbb{E}[f_1(x)g_1(y)] - \mathbb{E}[f(x)g(y)]| \leq 2\epsilon = \lambda\) as desired. Also, note that \(\hat{f}_1(\sigma) = \hat{f}(\sigma) \cdot \gamma^{||\sigma||}\) (similarly for \(\hat{g}_1(\sigma)\)). Thus, we get that,

\[\sum_{|\sigma| > d} \hat{f}_1(\sigma)^2 \leq \gamma^{2d} \cdot \sum_{|\sigma| > d} \hat{f}(\sigma)^2 \leq \gamma^{2d} \leq \eta,
\]

\[\sum_{|\sigma| > d} \hat{g}_1(\sigma)^2 \leq \gamma^{2d} \cdot \sum_{|\sigma| > d} \hat{g}(\sigma)^2 \leq \gamma^{2d} \leq \eta.
\]

\[\square\]

3.5 Joint Regularity Lemma for Fourier Concentrated Functions

The second step in our approach is to apply a regularity lemma on the functions \(f_1 : \mathcal{A}^n \to [-1, 1]\) and \(g_1 : \mathcal{B}^n \to [-1, 1]\) obtained from the previous step of smoothing. Regularity lemma is a loosely referred term which shows that for various types of combinatorial objects, an arbitrary object can be approximately decomposed into a constant number of “pseudorandom” sub-objects.

Our version of the regularity lemma draws inspiration from that of [DSTW10]; in fact, our proofs also closely follow theirs. Formally, we show the following lemma.

Lemma 3.5.1 (Joint Regularity Lemma for Fourier-Concentrated Functions). Let \((\mathcal{A} \times \mathcal{B}, \mu)\)
be a joint probability space. Let \( d \in \mathbb{N} \) and \( \tau > 0 \) be any given constant parameters. There exists an \( \eta \equiv \eta(\tau) > 0 \) and \( h \equiv h((A \times B, \mu), d, \tau) \) such that the following holds:

For all \( P \in L^2(A^n, \mu_A^n) \) and \( Q \in L^2(B^n, \mu_B^n) \) satisfying \( \sum_{|\sigma| > d} \hat{P}(\sigma)^2 \leq \eta \), \( \sum_{|\sigma| > d} \hat{Q}(\sigma)^2 \leq \eta \), and \( \text{Var}[P] \leq 1 \) and \( \text{Var}[Q] \leq 1 \), there exists a subset of indices \( H \subseteq [n] \) with \( |H| \leq h \), such that the restrictions of the functions \( P \) and \( Q \) obtained by evaluating the coordinates in \( H \) according to distribution \( \mu \) satisfy the following (where we denote \( T = [n] \setminus H \)):

- With probability at least \( 1 - \tau \) over \( \xi \sim \mu_A^{|H|} \), the restriction \( P_\xi(x_T) \) is such that for all \( i \in T \), it is the case that \( \text{Inf}_i(P_\xi(x_T)) \leq \tau \).

- With probability at least \( 1 - \tau \) over \( \xi \sim \mu_B^{|H|} \), the restriction \( Q_\xi(x_T) \) is such that for all \( i \in T \), it is the case that \( \text{Inf}_i(Q_\xi(x_T)) \leq \tau \).

In particular, one may take \( \eta = \tau^2/16 \) and \( h = \frac{d}{\tau^2} \cdot \left( \frac{C_4(\alpha)}{\alpha} \log \frac{C_4(\alpha)}{\alpha d^{\tau}} \right)^{O(d)} \) which is a constant that depends on \( d \), \( \tau \) and \( \alpha \equiv \alpha(\mu) \) (which is the minimum non-zero probability in \( \mu \)).

### 3.5.1 Regularity Lemma for Constant-Degree Polynomials

We first prove a version of the above regularity lemma for degree-\( d \) functions, as opposed to Fourier-concentrated functions.

**Lemma 3.5.2** (Regularity Lemma for Degree-\( d \) Functions). Let \((A, \mu_A)\) be a probability space. Let \( d \in \mathbb{N} \) and \( \tau > 0 \) be any given constant parameters. There exists \( h \equiv h((A, \mu_A), d, \tau) \) such that the following holds:

For all degree-\( d \) multilinear polynomials \( P \in L^2(A^n, \mu_A^n) \) with \( \text{Var}[P] \leq 1 \), there exists a subset of indices \( H_0 \subseteq [n] \) with \( |H_0| \leq h \), such that for any superset \( H \supseteq H_0 \), the restrictions of \( P \) obtained by evaluating the coordinates in \( H \) according to distribution \( \mu_A \) satisfy the following (where we denote \( T = [n] \setminus H \)):

\[
\Pr_{\xi \sim \mu_A^{|H|}} [\forall i \in T : \text{Inf}_i(P_\xi(x_T)) \leq \tau] \geq 1 - \tau.
\]

In other words, with probability at least \( 1 - \tau \) over the random restriction \( \xi \sim \mu_A^{|H|} \), the restricted function \( P_\xi(x_T) \) is such that \( \text{Inf}_i(P_\xi(x_T)) \leq \tau \) for all \( i \in T \).

In particular, one may take \( h = \frac{d}{\tau} \cdot \left( \frac{C_4(\alpha)}{\alpha} \log \frac{C_4(\alpha)}{\alpha d^{\tau}} \right)^{O(d)} \) which is a constant that depends on \( d \), \( \tau \) and \( \alpha \equiv \alpha(\mu_A) \).

The intuitive explanation of the regularity lemma is as follows: If \( P \) is a degree \( d \) polynomial with \( \text{Var}(P) \leq 1 \), then the total influence of \( P \) is at most \( d \). Hence, for all \( \beta > 0 \), there can only be at most \( h \equiv d/\beta \) variables with influence greater than \( \beta \). Indeed, our subset \( H_0 \) will essentially be the set of all the variables with influence at least \( \beta \) (we will choose \( \beta \) to be suitably smaller than \( \tau \), but with no dependence on \( n \)). Clearly, \( |H_0| \leq h \). For any
superset $H \supseteq H_0$, and for a random restriction of $x_H$ to $\xi$, it will follow from well-known hypercontractivity bounds (Theorem 3.2.14) and a careful union bound, that the influence of all the remaining variables will be less than $\tau$ with high probability.

Our regularity lemma draws inspiration from the one in [DSTW10]. In fact, our proof of the above regularity lemma also closely follows the proof steps in [DSTW10]. However, their regularity lemma was much more involved as they were dealing with low-degree polynomial threshold functions, whereas we are directly dealing with low-degree polynomials. In particular, a major difference in our regularity lemmas is that [DSTW10] obtain a (potentially) adaptive decision tree, whereas we obtain just a single subset $H$. Also, our notion of ‘regularity’ is much simpler in that we only need all influences to be small. Another aspect of our regularity lemma is that it is robust enough to also work for Fourier-concentrated functions, as opposed to only low-degree functions (potentially, [DSTW10] could also be modified to have this feature, although it was not required for their application). Another minor difference is that our Fourier analysis is for functions in $L^2(\mathcal{A}^n, \mu_\mathcal{A}^n)$, as opposed to functions on the boolean hypercube. But this is not really a significant difference and the proof steps go through as is, albeit with slightly different parameters which depend on the hypercontractivity parameters of the distribution $(\mathcal{A}, \mu_\mathcal{A})$.

Before we give a proof of Lemma 3.5.2, we need the following claim.

**Claim 3.5.3** (cf. Claim 3.12 in [DSTW10]). Let $P \in L^2(\mathcal{A}^n, \mu_\mathcal{A}^n)$ be a degree-$d$ polynomial. Let $H \subseteq [n]$ and $T = [n] \setminus H$. Let $\xi$ be a random restriction fixing $H$. For all $r \geq e^d$ and all $i \in T$, we have the following:

$$\Pr_{\xi} \left[\inf_i(P_\xi) > r \cdot C_4(\alpha)^d \cdot \inf_i(P)\right] \leq \exp(-c \cdot r^{1/d})$$

where $c = \alpha(\mu_\mathcal{A})d/e$ (see Theorem 3.2.14) and $C_4(\alpha)$ is obtained as in Theorem 3.2.13.

**Proof.** The identity $\inf_i(P_\xi) = \sum_{\sigma_T: (\sigma_T)_i \neq 0} \widehat{P}_\xi(\sigma_T)^2$ and Fact 3.2.12 imply that $\inf_i(P_\xi)$ is a degree-$2d$ polynomial in $\xi$. Hence, the claim would follow from the concentration bound for low-degree polynomials, i.e., Theorem 3.2.14, if we can appropriately upper-bound the $\ell_2$-norm of the polynomial $\inf_i(P_\xi)$. So, to prove Claim 3.5.3, it suffices to show that

$$\|\inf_i(P_\xi)\|_2 \leq C_4(\alpha)^d \cdot \inf_i(P). \quad (3.1)$$

By the triangle inequality for norms, we have that

$$\|\inf_i(P_\xi)\|_2 = \left\| \sum_{\sigma_T: (\sigma_T)_i \neq 0} \widehat{P}_\xi(\sigma_T)^2 \right\|_2 \leq \sum_{\sigma_T: (\sigma_T)_i \neq 0} \left\| \widehat{P}_\xi(\sigma_T)^2 \right\|_2.$$ 

Since $\widehat{P}_\xi(\sigma_T)$ is a degree-$d$ polynomial, the moment bound for low-degree polynomials, i.e., Theorem 3.2.13, yields that

$$\|\widehat{P}_\xi(\sigma_T)^2\|_2 = \|\widehat{P}_\xi(\sigma_T)\|_4^2 \leq C_4(\alpha)^d \left\| \widehat{P}_\xi(\sigma_T) \right\|_2^2.$$ 

82
and hence
\[
\|\text{Inf}_i(P_\xi)\|_2 \leq C_4(\alpha)^d \sum_{\sigma_T : (\sigma_T)_i \neq 0} \|\hat{P}_\xi(\sigma_T)\|_2^2 \\
= C_4(\alpha)^d \sum_{\sigma_T : (\sigma_T)_i \neq 0} \mathbb{E}_\xi \left[ \hat{P}_\xi(\sigma_T)^2 \right] \\
= C_4(\alpha)^d \cdot \mathbb{E}_\xi \left[ \text{Inf}_i(P_\xi) \right] \\
= C_4(\alpha)^d \cdot \text{Inf}_i(P)
\]

where the last equality follows from Lemma 3.2.11. Thus, Equation (3.1) and the claim follows from Theorem 3.2.14.

Proof of Lemma 3.5.2. Let \( P \in L^2(A^n, \mu_A^\otimes n) \) be the given degree-\( d \) multilinear polynomial with \( \text{Var}[P] \leq 1 \). From part (iii) of Fact 3.2.10, we have that \( \text{Inf}(P) \leq d \). Let \( H_0 \subseteq [n] \) be the set of indices \( i \in [n] \) such that \( \text{Inf}_i(f) \geq \beta \). Since \( d \geq \text{Inf}(P) \geq \sum_i \text{Inf}_i(P) \), we have that \( |H_0| \leq d/\beta \). We will choose \( \beta \) to be a suitable constant less than \( \tau \), but with no dependence on \( n \).

Fix \( H \supseteq H_0 \) and let \( T = [n] \setminus H \). From Claim 3.5.3, we have that for any \( i \in T \), it is the case that \( \text{Pr}_\xi \left[ \text{Inf}_i(P_\xi) > r \cdot C_4(\alpha)^d \cdot \text{Inf}_i(P) \right] \leq \exp(-\Omega(c \cdot r^{1/d})) \). However, to prove that \( \text{Inf}_i(P_\xi) \leq \tau \) for all \( i \in T \), with high probability, we cannot simply use a naive union bound over all \( i \in T \), as that will introduce a dependence of \( n \) in \( \beta \) and thereby in \( h \). Instead, we use a bucketing argument, as done in [DSTW10], as follows:

We partition the indices \( i \in T \) into buckets \( \{B_j\}_{j \in \mathbb{N}} \) as \( B_j = \{ i \in T : \text{Inf}_i(P) \in \left( \frac{\beta}{2j \tau}, \frac{\beta}{2j} \right] \} \). Since \( \text{Inf}(P) \leq d \), we have that \( |B_j| \leq 2^{j+1}d/\beta \). For all \( i \in B_j \), we use the concentration bound \( \text{Pr}_\xi \left[ \text{Inf}_i(P_\xi) \leq r \cdot C_4(\alpha)^d \cdot \text{Inf}_i(P) \right] \geq 1 - \exp(-c \cdot r^{1/d}) \) by choosing \( r = \frac{\tau 2^j}{\beta C_4(\alpha)^d} \). We then do a union bound over all the buckets. Thus, we get that:

\[
\text{Pr}_\xi \left[ \forall i \in T : \text{Inf}_i(P_\xi(x_T)) \leq \tau \right] \geq 1 - \sum_{j=0}^{\infty} \text{Pr}_\xi \left[ \exists i \in B_j : \text{Inf}_i(P_\xi(x_T)) > \tau \right] \\
\geq 1 - \sum_{j=0}^{\infty} \exp \left( -c \left( \frac{\tau \cdot 2^j}{\beta \cdot C_4(\alpha)^d} \right)^{1/d} \right) \cdot \frac{2^{j+1}d}{\beta}.
\]

It can be verified that for \( \frac{1}{\beta} = \frac{(2 \cdot C_4(\alpha))^d}{c^{d-\tau}} \cdot \log \left( \frac{(2 \cdot C_4(\alpha))^d}{c^{d-\tau}} \right)^d \) it holds that,

\[
\sum_{j=0}^{\infty} \exp \left( -c \left( \frac{\tau \cdot 2^j}{\beta \cdot C_4(\alpha)^d} \right)^{1/d} \right) \cdot \frac{2^{j+1}d}{\beta} \leq \tau.
\]

Thus, we have the regularity lemma as desired with \( |H_0| \leq h = \frac{d}{\beta} = \frac{d}{\tau} \left( \frac{C_4(\alpha)}{\alpha} \log \frac{C_4(\alpha)}{\alpha \cdot d \tau} \right)^{O(d)} \).
which is a constant that only depends on \(d\), \(\tau\) and \(\alpha \overset{\text{def}}{=} \alpha(\mu_A)\).

\[\] 

3.5.2 Joint Regularity Lemma

In this section, we use Lemma 3.5.2 to prove the joint regularity lemma, namely Lemma 3.5.1.

Proof of Lemma 3.5.1. We have \(P \in L^2(A^n, \mu_A^\otimes n)\) and \(Q \in L^2(B^n, \mu_B^\otimes n)\) satisfying the inequalities \(\sum_{|\sigma| > d} \hat{P}(\sigma)^2 \leq \eta, \sum_{|\sigma| > d} \hat{Q}(\sigma)^2 \leq \eta\), and \(\text{Var}[P] \leq 1, \text{Var}[Q] \leq 1\). First, we split \(P\) and \(Q\) into low and high degree components. That is, \(P(x) = P^\ell(x) + P^h(x)\) and \(Q(y) = Q^\ell(y) + Q^h(y)\), where \(P^\ell(x)\) and \(Q^\ell(y)\) contain all the monomials of degree at most \(d\) in \(P(x)\) and \(Q(y)\) respectively. Note that \(\text{Var}[P^\ell] \leq \text{Var}[P] \leq 1\). Similarly, \(\text{Var}[Q^\ell] \leq 1\).

We apply the regularity lemma for degree-\(d\) functions (Lemma 3.5.2), with parameter \(\tau\) equal to \(\tau/4\), on functions \(P^\ell\) and \(Q^\ell\) separately, to obtain subsets \(H_A, H_B \subseteq [n]\) respectively. The subset \(H\) is then obtained as \(H = H_A \cup H_B\). Note that \(|H| \leq h((A, \mu_A), d, \tau/4) + h((B, \mu_B), d, \tau/4)\), which is a computable function in terms of the parameters of the problem, but more importantly has no dependence on \(n\).

From Lemma 3.5.2, we know that for \(T = [n] \setminus H\) (note that \(H \supseteq H_A\) and \(H \supseteq H_B\)):

\[
\begin{align*}
\Pr_{\xi \sim \mu_A^{\otimes |H|}}[\forall i \in T : \text{Inf}_i(P^\ell_\xi(x_T)) \leq \tau/4] &\geq 1 - \tau/4, \quad (3.2) \\
\Pr_{\xi \sim \mu_B^{\otimes |H|}}[\forall i \in T : \text{Inf}_i(Q^\ell_\xi(y_T)) \leq \tau/4] &\geq 1 - \tau/4. \quad (3.3)
\end{align*}
\]

Now, we show that after adding \(P^h\) to \(P^\ell\), the influences \(\text{Inf}_i(P_\xi(x_T))\) are still upper-bounded by \(\tau\), with high probability over \(\xi\):

\[
\text{Inf}_i(P_\xi(x_T)) = \sum_{\sigma_T : |\sigma_T| \neq 0} \left( \sum_{\sigma_H} \hat{P}(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) \right)^2
\]

\[
= \sum_{\sigma_T : |\sigma_T| \neq 0} \left( \sum_{\sigma_H} \hat{P}^\ell(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) + \sum_{\sigma_H} \hat{P}^h(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) \right)^2
\]

\[
\leq 2 \cdot \sum_{\sigma_T : |\sigma_T| \neq 0} \left( \sum_{\sigma_H} \hat{P}^\ell(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) \right)^2 + \left( \sum_{\sigma_H} \hat{P}^h(\sigma_H \circ \sigma_T) \cdot \chi_{\sigma_H}(\xi) \right)^2
\]

\[
= 2 \cdot (\text{Inf}_i(P^\ell_\xi(x_T)) + \text{Inf}_i(P^h_\xi(x_T))). \quad (3.4)
\]

Since \(\E_\xi[\text{Var}(P^h_\xi(x_T))] \leq \text{Var}(P^h(x_T)) \leq \eta\) (see Lemma 3.2.11), we have by Markov’s inequality that

\[
\Pr_{\xi \sim \mu_A^{\otimes |H|}}[\text{Var}(P^h_\xi(x_T)) \leq 4\eta/\tau] \geq 1 - \tau/4.
\]

84
Since for all $i \in T$, we have $\inf_i (P^h_\xi(x_T)) \leq \text{Var}(P^h_\xi(x_T))$ (see Fact 3.2.10), we get that

$$\Pr_{\xi \sim \mu_\mathcal{A}^H} \left[ \forall i \in T : \inf_i (P^h_\xi(x_T)) \leq 4\eta / \tau \right] \geq 1 - \tau / 4. \quad (3.5)$$

We will choose $\eta = (\tau / 4)^2$, and thus, by a union bound (using Equations 3.4, 3.3 and 3.5), we have that

$$\Pr_{\xi \sim \mu_\mathcal{A}^H} \left[ \forall i \in T : \inf_i (P_\xi(x_T)) \leq \tau \right] \geq 1 - \tau / 2 > 1 - \tau.$$ 

By exactly the same sequence of calculations for $Q(y)$, we can have,

$$\Pr_{\xi \sim \mu_\mathcal{B}^H} \left[ \forall i \in T : \inf_i (Q_\xi(y_T)) \leq \tau \right] \geq 1 - \tau / 2 > 1 - \tau.$$

This completes the proof of Lemma 3.5.1.

\[ \square \]

### 3.6 Applying Correlation Bounds for Low-Influence Functions

The third step in our approach is to use correlation bounds for low-influence functions obtained from the invariance principle [MOO05, Mos10], to convert the functions $f_1 : \mathcal{A}^n \rightarrow [-1, 1]$ and $g_1 : \mathcal{B}^n \rightarrow [-1, 1]$ into functions $f_2 : \mathcal{A}^h \times \mathbb{R} \rightarrow [-1, 1]$ and $g_2 : \mathcal{B}^h \times \mathbb{R} \rightarrow [-1, 1]$ using the following lemma.

**Lemma 3.6.1 (Applying Correlation Bounds for Low-Influence Functions).** Let $(\mathcal{A} \times \mathcal{B}, \mu)$ be a joint probability space. Let $\gamma > 0$ be any given constant parameter. There exists a $\tau \defeq \tau((\mathcal{A} \times \mathcal{B}, \mu), \gamma) > 0$ such that the following holds:

For all functions $f_1 : \mathcal{A}^n \rightarrow [-1, 1]$ and $g_1 : \mathcal{B}^n \rightarrow [-1, 1]$, and a subset $H \subseteq [n]$ with $|H| = h$, such that the restrictions of the functions $f_1$ and $g_1$ obtained by evaluating the coordinates in $H$ according to distribution $\mu$, satisfy the following (where we denote $T = [n] \setminus H$):

- With probability at least $1 - \tau$ over $\xi \sim \mu_\mathcal{A}^H$, the restriction $(f_1)_\xi(x_T)$ is such that for all $i \in T$, it is the case that $\inf_i ((f_1)_\xi(x_T)) \leq \tau$.

- With probability at least $1 - \tau$ over $\xi \sim \mu_\mathcal{B}^H$, the restriction $(g_1)_\xi(x_T)$ is such that for all $i \in T$, it is the case that $\inf_i ((g_1)_\xi(x_T)) \leq \tau$.

There exist functions $f_2 : \mathcal{A}^h \times \mathbb{R} \rightarrow [-1, 1]$ and $g_2 : \mathcal{B}^h \times \mathbb{R} \rightarrow [-1, 1]$, such that,

$$\mathbb{E}_{x \sim \mu_\mathcal{A}^n} f_1(x) = \mathbb{E}_{x \sim \mu_\mathcal{A}^h, r_A \sim \mathcal{N}(0, 1)} f_2(x, r_A) \quad \text{and} \quad \mathbb{E}_{y \sim \mu_\mathcal{B}^n} g_1(y) = \mathbb{E}_{y \sim \mu_\mathcal{B}^h, r_B \sim \mathcal{N}(0, 1)} g_2(y, r_B)$$

85
and
\[
\mathbb{E}_{(x,y) \sim \mu \otimes h \sim \mathcal{G}(\rho)} [f_2(x, r_A) \cdot g_2(y, r_B)] \geq \mathbb{E}_{(x,y) \sim \mu \otimes \eta} [f_1(x) \cdot g_1(y)] - \gamma.
\]

Additionally, \(f_2\) and \(g_2\) will have the following special form: there exist functions \(f'_2: \mathcal{A} \to \mathbb{R}\) and \(g'_2: \mathcal{B} \to \mathbb{R}\) such that,
\[
f_2(x, r) = \begin{cases} 1 & r \geq f'_2(x) \\ -1 & r < f'_2(x) \end{cases}
\]
and
\[
g_2(y, r) = \begin{cases} 1 & r \geq g'_2(y) \\ -1 & r < g'_2(y) \end{cases}.
\]
Also, one may take \(\tau = \gamma^{O\left(\frac{\log(1/\gamma) \log(1/\alpha)}{(1-\rho)^2}\right)}\) where \(\rho = \rho(\mathcal{A}, \mathcal{B}; \mu)\) and \(\alpha \equiv \alpha(\mu)\) (which is the minimum non-zero probability in \(\mu\)).

As mentioned before, the main technical tool in proving Lemma 3.6.1 is a result about correlation bounds for low-influence functions (which are generalizations of the 'Majority is Stablest' theorem). Before we state that theorem, we need the following definition, which is a slightly modified version of Definition 1.12 in [Mos10].

**Definition 3.6.2 (Gaussian Stability).** Let \(\Phi\) be the cumulative distribution function (CDF) of a standard \(\mathcal{N}(0,1)\) Gaussian. Given \(\rho \in [-1,1]\) and \(\mu, \nu \in [-1,1]\), we define,
\[
\Gamma_{\rho}(\mu, \nu) = \mathbb{E}[\overline{P}_{\mu}(X) \cdot \overline{Q}_{\nu}(Y)]
\]
\[
\Gamma_{\rho}(\mu, \nu) = -\mathbb{E}[\overline{P}_{\mu}(X) \cdot \overline{Q}_{\nu}(Y)]
\]
where \((X, Y)\) is distributed according to \(\mathcal{G}(\rho)\) and
\[
\overline{P}_{\mu}(X) = \begin{cases} 1 & X \leq \Phi^{-1}(\frac{1+\mu}{2}) \\ -1 & \text{otherwise} \end{cases}
\]
and \(\overline{Q}_{\nu}(X) = \begin{cases} 1 & Y \leq \Phi^{-1}(\frac{1+\nu}{2}) \\ -1 & \text{otherwise} \end{cases}\).

Note that for \((X, Y) \sim \mathcal{G}(\rho)\), we have that
\[
\mathbb{E}_X [\overline{P}_{\mu}(X)] = \mu \quad \text{and} \quad \mathbb{E}_Y [\overline{Q}_{\nu}(Y)] = \nu = \mathbb{E}_Y [-\overline{Q}_{-\nu}(Y)].
\]

With this definition in hand, we can state the correlation bounds for low-influential functions that are obtained from the invariance principle.

**Theorem 3.6.3 (Correlation Bounds from Invariance Principle; [MOO05, Mos10]).** Let \((\mathcal{A} \times \mathcal{B}, \mu)\) be a joint probability space. As before, let \(\alpha = \alpha(\mu)\) be the minimum probability of any atom in \(\mathcal{A} \times \mathcal{B}\). Let \(\rho = \rho(\mathcal{A}, \mathcal{B}; \mu)\) be the maximal correlation of the joint probability space (see Definition 3.2.16).

Then, for all \(\epsilon > 0\), there exists \(\tau \equiv \tau((\mathcal{A} \times \mathcal{B}, \mu), \epsilon) > 0\) such that if
\[
P : \mathcal{A}^n \to [-1,1] \quad \text{and} \quad Q : \mathcal{B}^n \to [-1,1]
\]

satisfy \( \text{Inf}_i(P) \leq \tau \) and \( \text{Inf}_i(Q) \leq \tau \) for all \( i \in [n] \), then
\[
\Gamma_\rho (\mathbb{E}_x[P(x)] , \mathbb{E}_y[Q(y)]) - \epsilon \leq \mathbb{E}_{(x,y) \sim \mu^\otimes n} [P(x)Q(y)] \leq \Gamma_\rho (\mathbb{E}_x[P(x)] , \mathbb{E}_y[Q(y)]) + \epsilon
\]
Furthermore, one may take
\[
\tau = \epsilon O \left( \frac{\log(1/\epsilon) \log(1/\alpha)}{(1-\rho)\epsilon^2} \right).
\]

Intuitively, this theorem says that if \( P \) and \( Q \) are low-influential, then their correlation is not much more than that of appropriate threshold functions applied on \( \rho \)-correlated Gaussians. With this tool in hand, we are now ready to prove Lemma 3.6.1.

**Proof of Lemma 3.6.1.** Suppose we have \( f_1 : A^n \rightarrow [-1,1] \) and \( g_1 : B^n \rightarrow [-1,1] \), and a subset \( H \subseteq [n] \) with \( |H| = h \), such that the restrictions of the functions \( f_1 \) and \( g_1 \) obtained by evaluating the coordinates in \( H \) according to distribution \( \mu \), satisfy the properties as stated in the lemma. We construct function \( f_2 : A^h \times \mathbb{R} \rightarrow [-1,1] \) and \( g_2 : B^h \times \mathbb{R} \rightarrow [-1,1] \) by replacing the functions obtained after restricting the variables in \( H \) by appropriate threshold functions acting on \( \rho \)-correlated Gaussians, namely:
\[
\forall (x, r) \in A^h \times \mathbb{R} : f_2(x, r) = \overline{P}_{\nu_1}(r) \text{ where } \nu_1 \overset{\text{def}}{=} \mathbb{E}_{x_T \sim \mu_{A}^{\otimes n-h}} [f_1(x_H \leftarrow x, x_T)],
\]
\[
\forall (y, r) \in B^h \times \mathbb{R} : g_2(y, r) = \overline{Q}_{\nu_2}(r) \text{ where } \nu_2 \overset{\text{def}}{=} \mathbb{E}_{y_T \sim \mu_{B}^{\otimes n-h}} [f_1(y_H \leftarrow y, y_T)],
\]
where \( \overline{P}_{\nu_1} \) and \( \overline{Q}_{\nu_2} \) are as defined in Definition 3.6.2.\(^7\)

It follows by definition, that \( \mathbb{E}[f_2(x, r)] = \mathbb{E}[f_1(x)] \) and \( \mathbb{E}[g_2(y, r)] = \mathbb{E}[g_1(y)] \). That is, this process has not changed the individual means of \( f_1 \) and \( g_1 \). We now need to prove that the correlation is not hurt by much. From Lemma 3.5.1 and a simple union bound, we know that with probability \( 1 - 2\tau \), a random restriction \((x_H, y_H)\) for the coordinates in \( H \) is such that:
\[
\forall i \in T : \text{Inf}_i((f_1)_x(x_T)) \leq \tau \text{ and } \text{Inf}_i((g_1)_y(y_T)) \leq \tau.
\]
Let us call all the tuples \((x_H, y_H)\) for which the above happens as ‘good’. Then, we have that:

\(^7\)For simplicity, we will abuse notations in the following sense: when we say \( f_1(x) \), we mean \( x \in A^n \), but when we say \( f_2(x, r) \), we mean \( x \in A^h \) and \( r \in \mathbb{R} \).
$E_{x,y}f_1(x)g_1(y)$

$= E_{x_H,y_H}[E_{x_T,y_T}f_1(x_H,x_T) \cdot g_1(y_H,y_T)]$

$= \Pr[(x_H,y_H) \text{ is not 'good']} \cdot E_{x_H,y_H}[E_{x_T,y_T}f_1(x_H,x_T) \cdot g_1(y_H,y_T)|(x_H,y_H) \text{ is not 'good'}]$

$+ \Pr[(x_H,y_H) \text{ is 'good']} \cdot E_{x_H,y_H}[E_{x_T,y_T}f_1(x_H,x_T) \cdot g_1(y_H,y_T)|(x_H,y_H) \text{ is 'good'}]$

$\leq \Pr[(x_H,y_H) \text{ is not 'good'}] \cdot 1$

$+ \Pr[(x_H,y_H) \text{ is 'good']} \cdot E_{x_H,y_H}[E_{r_A,r_B}f_2(x_H,r_A) \cdot g_2(y_H,r_B) + \epsilon \cdot (x_H,y_H) \text{ is 'good'}]$

$= \Pr[(x_H,y_H) \text{ is not 'good']} \cdot \left(1 - E_{x_H,y_H}[E_{r_A,r_B}f_2(x_H,r_A) \cdot g_2(y_H,r_B) + \epsilon \cdot (x_H,y_H) \text{ is not 'good'}]\right)$

$+ E_{x_H,y_H}[E_{r_A,r_B}f_2(x_H,r_A) \cdot g_2(y_H,r_B) + \epsilon]$

$\leq E_{x_H,y_H}[E_{r_A,r_B}f_2(x_H,r_A) \cdot g_2(y_H,r_B)] + 2\tau \cdot (2 - \epsilon) + \epsilon$

$\leq E_{x_H,y_H}[E_{r_A,r_B}f_2(x_H,r_A) \cdot g_2(y_H,r_B)] + 2\epsilon.$

Step 3 above is due to the definition of $f_2$ and $g_2$ and Theorem 3.6.3. The last step follows because $\tau \ll \epsilon$, and so we can upper bound $2\tau \cdot (2 - \epsilon) \leq \epsilon$.

Thus, finally we choose $\epsilon = \gamma/2$ for Theorem 3.6.3, and we get $\tau = \tau(\gamma)$ accordingly, thereby getting the final requirement of Lemma 3.6.1, that is,

$$E_{(x_H,y_H) \sim \mu^{\otimes h}}f_2(x_H,r_A) \cdot g_2(y_H,r_B) \geq E_{(x,y) \sim \mu^{\otimes n}}f_1(x) \cdot g_1(y) - \gamma.$$ 

\[\square\]

### 3.7 Simulating Correlated Gaussians

In this section, we use a technique due to Witsenhausen [Wit75] which shows that for any joint probability space $(A \times B, \mu)$ with maximal correlation $\rho$, Alice and Bob can non-interactively simulate $\rho$-correlated Gaussians up to arbitrarily small two-dimensional Kolmogorov distance. We obtain the following lemma.

**Lemma 3.7.1 (Witsenhausen’s Rounding).** Let $(\mathcal{A} \times \mathcal{B}, \mu)$ be a joint probability space, and let $\rho = \rho(\mathcal{A}, \mathcal{B}; \mu)$ be its maximal correlation. Let $\zeta > 0$ be any given parameter. Then, there exists $w \overset{\text{def}}{=} w((\mathcal{A} \times \mathcal{B}, \mu), \zeta) \in \mathbb{N}$ such that the following holds:

Let $f_2 : \mathcal{A}^h \times \mathbb{R} \rightarrow [-1,1]$ and $g_2 : \mathcal{B}^h \times \mathbb{R} \rightarrow [-1,1]$ be functions for which there exist functions $f'_2 : \mathcal{A}^h \rightarrow \mathbb{R}$ and $g'_2 : \mathcal{B}^h \rightarrow \mathbb{R}$ such that
\[
\begin{align*}
    f_2(x, r) &= \begin{cases} 1 & r \geq f'_2(x) \\ -1 & r < f'_2(x) \end{cases} \quad \text{and} \quad g_2(y, r) = \begin{cases} 1 & r \geq g'_2(y) \\ -1 & r < g'_2(y) \end{cases},
\end{align*}
\]

Then, there exist functions \( f_3 : \mathcal{A}^{h+w} \to [-1, 1] \) and \( g_3 : \mathcal{B}^{h+w} \to [-1, 1] \) such that

\[
\begin{align*}
    \left| \mathbb{E}_{x \sim \mu_A^{\otimes (h+w)}} f_3(x) - \mathbb{E}_{x \sim \mu_A^{\otimes h}} f_2(x, r_A) \right| &\leq \zeta, \\
    \left| \mathbb{E}_{y \sim \mu_B^{\otimes (h+w)}} g_3(y) - \mathbb{E}_{y \sim \mu_B^{\otimes h}} g_2(y, r_B) \right| &\leq \zeta,
\end{align*}
\]

and

\[
\left| \mathbb{E}_{(x,y) \sim \mu^{\otimes (h+w)}} \left[ f_3(x) \cdot g_3(y) - \mathbb{E}_{(x,y) \sim \mu^{\otimes h}} \left[ f_2(x, r_A) \cdot g_2(y, r_B) \right] \right] \right| \leq \zeta.
\]

In particular, one may take \( w = O \left( \frac{1+\rho}{\alpha(1-\rho)\zeta^2} \right) \), where \( \alpha \overset{\text{def}}{=} \alpha(\mu) \) is the minimum non-zero probability in \( \mu \).

The main idea in obtaining the functions \( f_3 \) and \( g_3 \) is the technique of Witsenhausen [Wit75] for simulating \( \rho \)-correlated Gaussians from many copies of \( (\mathcal{A} \times \mathcal{B}, \mu) \).

**Lemma 3.7.2 (Simulating Gaussians [Wit75]).** Let \( (\mathcal{A} \times \mathcal{B}, \mu) \) be a joint probability space, and let \( \rho = \rho(\mathcal{A}, \mathcal{B}; \mu) \) be its maximal correlation. Let \( \zeta > 0 \) be any given parameter. Then, there exists \( w \overset{\text{def}}{=} w((\mathcal{A} \times \mathcal{B}, \mu), \zeta) \in \mathbb{N} \) such that the following holds:

For all \( \nu_1, \nu_2 \in [-1, +1] \), there exist functions \( P_{\nu_1} : \mathcal{A}^w \to [-1, 1] \) and \( Q_{\nu_2} : \mathcal{B}^w \to [-1, 1] \) such that \( |\mathbb{E}[P_{\nu_1}(x)] - \nu_1| \leq \zeta/2 \), \( |\mathbb{E}[Q_{\nu_2}(y)] - \nu_2| \leq \zeta/2 \) and

\[
\left| \mathbb{E}_{(x,y) \sim \mu^{\otimes w}} [P_{\nu_1}(x)Q_{\nu_2}(y)] - \Gamma_\rho(\nu_1, \nu_2) \right| \leq \zeta.
\]

In particular, one may take \( w = O \left( \frac{1+\rho}{\alpha(1-\rho)\zeta^2} \right) \) where \( \alpha \overset{\text{def}}{=} \alpha(\mu) \).

**Proof.** Since \( \rho = \rho(\mathcal{A}, \mathcal{B}; \mu) \), we have (by the definition of maximal correlation) that there exist functions \( f : \mathcal{A} \to \mathbb{R} \) and \( g : \mathcal{B} \to \mathbb{R} \) such that \( \mathbb{E}_{x \sim \mu_A} f(x) = \mathbb{E}_{y \sim \mu_B} g(y) = 0 \), \( \text{Var}(f) = 1 \) and \( \mathbb{E}_{(x,y) \sim \mu} [f(x) \cdot g(y)] = \rho \).

We define \( F(x) = \frac{\sum_{i=1}^{w} f(x_i)}{\sqrt{w}} \) and \( G(y) = \frac{\sum_{i=1}^{w} g(y_i)}{\sqrt{w}} \), and define \( P_{\nu_1} \) and \( Q_{\nu_2} \) as follows:

\[
P_{\nu_1}(x) = \begin{cases} 1 & F(x) \leq \Phi^{-1}(\frac{1+\nu_1}{2}) \\ -1 & \text{otherwise} \end{cases} \quad \text{and} \quad Q_{\nu_2}(y) = \begin{cases} 1 & G(y) \leq \Phi^{-1}(\frac{1+\nu_2}{2}) \\ -1 & \text{otherwise} \end{cases}.
\]

We apply Lemma 3.2.22 for the pair of random variables \((f(x), g(y))\) with parameter \( \zeta \) being \( \zeta/4 \), to obtain the appropriate \( w \). It easily follows that, \( |\mathbb{E}[P_{\nu_1}(x)] - \nu_1| \leq \zeta/2 \) and \( |\mathbb{E}[Q_{\nu_2}(y)] - \nu_2| \leq \zeta/2 \) and

\[
\left| \mathbb{E}_{(x,y) \sim \mu^{\otimes w}} [P_{\nu_1}(x)Q_{\nu_2}(y)] - \Gamma_\rho(\nu_1, \nu_2) \right| \leq \zeta.
\]

\( \square \)
We are now ready to prove Lemma 3.7.1.

**Proof of Lemma 3.7.1.** Given \((A \times B, \mu)\) and \(\zeta\), we obtain \(w\) as in Lemma 3.7.2. Given functions \(f_2\) and \(g_2\), of the said form, we construct functions \(f_3 : A^{h+w} \rightarrow [-1, 1]\) and \(g_3 : B^{h+w} \rightarrow [-1, 1]\) by invoking Lemma 3.7.2 for every assignment to the first \(h\) variables with parameter \(\zeta\). In particular, for every \(x_1 \in A^h, x_2 \in A^w\), we define \(f_3(x_1, x_2) = P_{f_2(x_1)}(x_2)\). Similarly, for \(y_1 \in B^h, y_2 \in A^w\), we define \(g_3(y_1, y_2) = Q_{g_2(y_1)}(y_2)\). This gives us that \(|\mathbb{E}[f_3(x)] - \mathbb{E}[f_2(x, r_A)]| \leq \zeta/2\), \(|\mathbb{E}[g_3(y)] - \mathbb{E}[g_2(y, r_B)]| \leq \zeta/2\) and

\[
\left| \mathbb{E}_{(x,y) \sim \mu \otimes (h+w)} [f_3(x) \cdot g_3(y)] - \mathbb{E}_{(x,y) \sim \mu \otimes h} [f_2(x, r_A) \cdot g_2(y, r_B)] \right| \leq \zeta.
\]

Thus, we have \(f_3\) and \(g_3\) as desired. \(\square\)

### 3.8 Putting it All Together!

In this section, we finally use all the lemmas we have developed to prove Theorem 3.3.1.

**Proof of Theorem 3.3.1.** Given \((A \times B, \mu)\) and \(\delta > 0\) and functions \(f : A^n \rightarrow [-1, 1]\) and \(g : B^n \rightarrow [-1, 1]\), we wish to apply Lemma 3.6.1 with parameter \(\gamma = \delta/3\) followed by Lemma 3.7.1 with parameter \(\zeta = \delta/3\). Lemma 3.6.1 will dictate a value \(\tau = \tau((A \times B, \mu), \gamma)\). We wish to apply the joint regularity lemma (Lemma 3.5.1) with this parameter \(\tau\), which will dictate a value of \(\eta = \eta(\tau)\). Using this value of \(\eta\), and \(\lambda = \delta/3\), we apply the smoothing lemma (Lemma 3.4.1), which will dictate a value of \(d = d((A \times B, \mu), \lambda, \eta)\). We use this \(d\) to feed into the joint regularity lemma (Lemma 3.5.1) and to obtain a value of \(h\). The final value of \(n_0\) is the sum of \(h((A \times B, \mu), d, \tau)\) given by the joint regularity lemma (Lemma 3.5.1) and \(w((A \times B, \mu), \zeta)\) given by Witsenhausen’s rounding procedure (Lemma 3.7.1). This dependency of parameters is pictorially described in Figure 3-3 (the dependencies on \((A \times B, \mu)\) are suppressed for the sake of clarity). It can be shown by putting everything together that \(n_0 = \exp \left( \text{poly} \left( \frac{1}{\delta}, \frac{1}{1-\eta}, \log \left( \frac{1}{\lambda} \right) \right) \right)\).

Once we have all the parameters set, we are now able to apply them to any pair of functions \(f : A^n \rightarrow [-1, 1]\) and \(g : B^n \rightarrow [-1, 1]\). In particular, we proceed as described in the overview (Section 3.3).

**Step 1:** We apply Lemma 3.4.1 to functions \(f\) and \(g\) with parameters \(\lambda\) and \(\eta\) as obtained above. This gives us a degree \(d\) and functions \(f_1\) and \(g_1\), such that \(\sum_{|\sigma| > d} \tilde{g}(\sigma)^2 < \eta\) and \(\sum_{|\sigma| > d} \tilde{f}(\sigma)^2 < \eta\).

**Step 2:** We apply the joint regularity lemma (Lemma 3.5.1) on functions \(f_1\) and \(g_1\), with parameters \(d\) and \(\tau\) as obtained above (note that the conditions involving \(\eta\) are satisfied because we chose precisely this \(\eta\) to be given to the smoothing lemma). This gives us a subset \(H \subseteq [n]\) such that \(|H| \leq h\) and with high probability over restrictions to this
subset $H$, the restricted versions of both $f_1$ and $g_1$ have all individual influences to be at most $\tau$.

**Step 3:** We apply the correlation bounds result (Lemma 3.6.1) to functions $f_1$ and $g_1$ (note that all the conditions involving $\tau$ are satisfied already because we chose precisely this $\tau$ to be given to the joint regularity lemma).

This gives us functions $f_2 : \mathcal{A}^h \times \mathbb{R} \to [-1, 1]$ and $g_2 : \mathcal{B}^h \times \mathbb{R} \to [-1, 1]$ of the form: there exist functions $f'_2 : \mathcal{A}^h \to \mathbb{R}$ and $g'_2 : \mathcal{B}^h \to \mathbb{R}$ such that

$$f_2(x, r) = \begin{cases} 1 & r \geq f'_2(x) \\ -1 & r < f'_2(x) \end{cases} \quad \text{and} \quad g_2(y, r) = \begin{cases} 1 & r \geq g'_2(y) \\ -1 & r < g'_2(y) \end{cases}.$$ 

**Step 4:** Functions $f_2$ and $g_2$ are exactly in the form for which Lemma 3.7.1 is applicable, which we use with parameters $\zeta$ as obtained above. This gives us functions $f_3 : \mathcal{A}^{h+w} \to [-1, 1]$ and $g_3 : \mathcal{B}^{h+w} \to [-1, 1]$.

Note that $\mathbb{E}f = \mathbb{E}f_1 = \mathbb{E}f_2$ and $|\mathbb{E}f_3 - \mathbb{E}f_2| \leq \zeta = \delta/3$ and similarly $\mathbb{E}g = \mathbb{E}g_1 = \mathbb{E}g_2$ and $|\mathbb{E}g_3 - \mathbb{E}g_2| \leq \zeta = \delta/3$. Moreover, we have from Lemmas 3.7.1, 3.6.1 and 3.4.1 that

$$\mathbb{E}_{(x,y) \sim \mu^{(h+w)}} [f_3(x) \cdot g_3(y)] \geq \mathbb{E}_{(r,x) \sim \mu^h} [f_2(x) \cdot g_2(y)] - \zeta \geq \mathbb{E}_{(x,y) \sim \mu^n} [f_1(x) \cdot g_1(y)] - \gamma - \zeta \geq \mathbb{E}_{(x,y) \sim \mu^n} [f(x) \cdot g(y)] - \lambda - \gamma - \zeta = \mathbb{E}_{(x,y) \sim \mu^n} [f(x) \cdot g(y)] - \delta$$

Hence, taking $\tilde{f} = f_3$ and $\tilde{g} = g_3$, proves Theorem 3.3.1.
3.8.1 Generalizing to Arbitrary Binary Targets

We now give a proof sketch of Theorem 3.2.3. Even though this is not a black-box application of Theorem 7.1.3, it follows the same proof steps. We highlight the main differences in this section.

We consider two cases, (I) \( \mathbb{E}[UV] \geq \mathbb{E}[U] \cdot \mathbb{E}[V] \) and (II) \( \mathbb{E}[UV] \leq \mathbb{E}[U] \cdot \mathbb{E}[V] \).

**Case (I):** \( \mathbb{E}[UV] \geq \mathbb{E}[U] \cdot \mathbb{E}[V] \)

We need to modify the GAP-BAL-MAX-INNER-PRODUCT problem (Problem 3.2.6), by replacing the conditions on \( \| E[f(x)] \| \) by \( \| E[f(x)] - \mathbb{E}[U] \| \), and similarly replacing the conditions on \( \| E[g(y)] \| \) by \( \| E[g(y)] - \mathbb{E}[V] \| \) and replacing \( \rho \) by \( \mathbb{E}[UV] \). The reduction between GAP-NON-INT-SIM and GAP-BAL-MAX-INNER-PRODUCT works in almost exactly the same way.

It is easy to see that using the main technical theorem (Theorem 3.3.1) and following the same proof as of Theorem 7.1.3, we also get decidability for GAP-NON-INT-SIM((\( \mathcal{A} \times \mathcal{B}, \mu \)), (\( \mathcal{U} \times \mathcal{V}, \nu \)), \( \delta \)).

**Case (II):** \( \mathbb{E}[UV] \leq \mathbb{E}[U] \cdot \mathbb{E}[V] \)

As in the previous case, we need to modify the GAP-BAL-MAX-INNER-PRODUCT problem (Problem 3.2.6), by replacing the conditions on \( \| E[f(x)] \| \) by \( \| E[f(x)] - \mathbb{E}[U] \| \), and similarly replacing the conditions on \( \| E[g(y)] \| \) by \( \| E[g(y)] - \mathbb{E}[V] \| \) and replacing \( \rho \) by \( \mathbb{E}[UV] \). The condition on \( E[f(x)g(y)] \) will however change as \( E[f(x)g(y)] \leq E[UV] + \delta \) in case (i) versus. \( E[f(x)g(y)] \geq E[UV] + 4\delta \) in case (ii). The reduction between GAP-NON-INT-SIM and GAP-BAL-MAX-INNER-PRODUCT works in almost exactly the same way.

The main difference in this case however is that, we want each of the steps to ‘increase’ correlation by a small amount as opposed to ‘decrease’ the correlation. In particular, the main condition in Theorem 3.3.1 will change as follows:

\[
\mathbb{E}_{(x,y) \sim \mu \otimes n_0} \left[ \tilde{f}(x) \cdot \tilde{g}(y) \right] \leq \mathbb{E}_{(x,y) \sim \mu \otimes n} [f(x) \cdot g(y)] + \delta.
\]

The steps of smoothing (Lemma 3.4.1), joint regularity (Lemma 3.5.1) and Witsenhausen rounding (Lemma 3.7.1) do not need any modification as they approximately preserve the correlation in both directions. However, in the step of applying Correlation Bounds (Lemma 3.6.1), we need to use the lower bound of \( \Gamma_{\rho}(\cdot, \cdot) \) instead of the upper bound of \( \overline{\Gamma}_{\rho}(\cdot, \cdot) \). In particular, the lemma will change slightly resulting in functions such that,

\[
\mathbb{E}_{(x,y) \sim \mu \otimes h, (r_A, r_B) \sim \mathbb{G}(\rho)} [f_2(x, r_A) \cdot g_2(x, r_B)] \leq \mathbb{E}_{(x,y) \sim \mu \otimes n} [f_1(x) \cdot g_1(y)] + \gamma.
\]
Additionally, \( f_2 \) and \( g_2 \) will have the following special form: there exist functions \( f'_2 : \mathcal{A}^h \to \mathbb{R} \) and \( g'_2 : \mathcal{B}^h \to \mathbb{R} \) such that

\[
f_2(x, r) = \begin{cases} 1 & r \geq f'_2(x) \\ -1 & r < f'_2(x) \end{cases}
\quad \text{and} \quad
g_2(y, r) = \begin{cases} -1 & r \geq g'_2(y) \\ 1 & r < g'_2(y) \end{cases}.
\]

This structural difference in \( f_2 \) and \( g_2 \) affects the Witsenhausen rounding step (Lemma 3.7.1) slightly, but it is easy to see that the same proof strategy works.

It is also easy to see that using this modified main theorem (analogue of Theorem 3.3.1) and following the same proof steps as of Theorem 7.1.3, we also get decidability for \( \text{Gap-Non-Int-Sim}(\mathcal{A} \times \mathcal{B}, \mu), (\mathcal{U} \times \mathcal{V}, \nu), \delta) \) in this case.

### 3.9 Open Questions

In this chapter, we proved computable bounds on the non-interactive simulation of any \( 2 \times 2 \) distribution. We now conclude with some interesting open questions.

The running time of our algorithm is at least doubly-exponential in the input size\(^8\). It would be very interesting to understand the computational complexity of the non-interactive simulation problem. We point out that the question of generating the best DSBS can be thought of as a tensored version of the following “\text{Min-Bipartite-Bisection}” problem: We are given a weighted bipartite graph \( G = (L \cup R, E) \), and we wish to find a subset \( S \) of \( L \cup R \) such that \( S \cap L \) roughly contains half the vertices of \( L \), and \( S \cap R \) roughly contains half the vertices of \( R \), while minimizing the total weight of edges crossing the cut \( (S, \overline{S}) \). While it follows from [RST12] that \text{Min-Bipartite-Bisection} is hard to approximate, the same is not necessarily true about its tensored version.

Another interesting open question is to generalize our computability results to more than two players, which also seems to require new technical ideas.

Finally, it would be very interesting to see if our techniques could apply to other “tensored” problems. The most relevant problems seem to be (i) computing the zero-error Shannon capacity of a graph [Sha56, Lov79, AL06], (ii) deciding a quantum version of our problem, namely that of local state transformation of quantum entanglement [Bei12, DB13] and (ii) approximately computing the entangled value of a \( 2 \)-prover \( 1 \)-round game ([KKM+11]; also see the open problem [ope]).

---

\(^8\)For constant values of \( \delta \) and \( \rho \), the running time is doubly-exponential in \( 2^{\text{poly} \log m} \). Here, we think of the input as a bipartite graph with \( m \) edges. This follows because \( \alpha \sim 1/m \).
Chapter 4

Communication with Imperfectly Shared Randomness

4.1 Introduction & Related Work

In this chapter, we give an essentially tight procedure for converting a two-way public-coin protocol into a protocol in the simultaneous message passing model using imperfectly shared randomness.

In the communication with imperfectly randomness framework, which was first studied by [BGI14, CGMS14] (see also [GKS16a] and [GS17]), Alice and Bob wish to compute a joint function of their inputs and have access to i.i.d. samples from a known source. The most basic and natural source of binary correlation is given by $\rho$-correlated bits, which denote by $\text{DSBS}(\rho)$ (the doubly symmetric binary source with correlation parameter $\rho$). Namely, for any $-1 \leq \rho \leq 1$, we say that $(X, Y) \sim \text{DSBS}(\rho)$ if $X, Y$ are both uniform over $\{-1, 1\}$ and their correlation (and covariance) is $E[XY] = \rho$ (i.e., this source corresponds to a binary symmetric channel with uniform input).

With $\text{DSBS}(\rho)$ as the source of correlation (for instance), the setup of communication with imperfectly shared randomness interpolates between the well-studied public randomness ($\rho = 1$) and private randomness ($\rho = 0$) models. We also point out that communication with imperfectly shared randomness is closely related to common randomness generation, studied in Chapter 2. For instance, an efficient common randomness generation protocol leads to a simple approach for communication with imperfectly shared randomness: first generate common randomness and then use it to run an efficient public-coin protocol. While this approach can lead to some savings in communication, it turns out that in several cases, one can substantially reduce the communication without first solving the common randomness generation problem.

Recall that in the simultaneous message passing (henceforth denoted by SMP) model, each of Alice and Bob can send a single message to the referee who should then output the answer of the protocol (see Figure 4-1).

In the SMP model, [BGI14] exhibit a (partial) function whose communication complexity
using $\text{DSBS}(\rho)$ (as the source of imperfectly shared randomness) is exponentially larger than the SMP communication complexity using public randomness (for any constant $\rho < 1$). In this chapter, we prove that this separation is tight. In fact, we show a stronger result that every function having two-way (i.e., interactive) communication $c$ bits using public randomness has a SMP protocol with $2^{O(c)}$ bits using $\text{DSBS}(\rho)$ for every constant $\rho < 1$. This answers a question of Sudan [Sud14]. Moreover, we are able to prove a more general result that applies to a broad family of sources of correlated randomness including $\text{DSBS}(\rho)$ (as well as $\rho$-correlated Gaussians):

**Theorem 4.1.1.** Let $\rho \in (0, 1]$ and $\mu$ be any source of randomness with maximal correlation $\rho$. Every (possibly partial) function $f$ with $(1/3)$-error two-way communication $c$ bits with perfectly shared randomness has $\delta$-error SMP communication with $\mu$-randomness at most $2^{O(c)} \cdot \log(1/\delta)/\rho^2$ bits for every $\delta > 0$.

We point out that the above theorem applies to $\text{DSBS}(\rho)$ as a special case because of the fact (due to [Wit75]) that the maximal correlation of $\text{DSBS}(\rho)$ is equal to $\rho$.

We prove Theorem 4.1.1 in the next section.

### 4.2 Proof of Theorem 4.1.1

In order to prove Theorem 4.1.1, we will give a simultaneous message passing (SMP) protocol with $\mu$-randomness (where $\mu$ is any source of randomness with maximal correlation $\rho$) solving the following problem which is equivalent to “sketching $\ell_2$-norms on the unit sphere.” This problem was studied by [CGMS14] to prove a one-way (instead of an SMP) analogue of Theorem 4.1.1.
Definition 4.2.1 (GapInnerProduct_{r,s}). Let \(-1 \leq s < r \leq 1\) be known to Alice and Bob. Alice is also given a unit vector \(u \in \mathbb{R}^n\) and Bob is given a unit vector \(v \in \mathbb{R}^n\). The goal is for Alice and Bob to distinguish the case where \(\langle u, v \rangle \geq r\) from the case where \(\langle u, v \rangle \leq s\).

The next lemma shows that GapInnerProduct is complete for functions with low interactive communication complexity.

Lemma 4.2.2 ([CGMS14]). Let \(f\) be a (possibly partial) two-party function \(f : \{0,1\}^{2n} \rightarrow \{0,1\}\) such that \(f\) has \((1/3)\)-error two-way communication complexity \(c\) bits with perfect randomness. Then, there exists a function \(\ell(n) \in \mathbb{N}\) along with mappings \(g_A : \{0,1\}^n \rightarrow \{\pm \frac{1}{\sqrt{\ell(n)}}\}^{\ell(n)}\) and \(g_B : \{0,1\}^n \rightarrow \{\pm \frac{1}{\sqrt{\ell(n)}}\}^{\ell(n)}\) such that

- If \(f(x,y) = 0\), then \((g_A(x), g_B(y))\) is a NO instance of \(\text{GapInnerProduct}_{\frac{1}{3} \cdot 2^{-k} - 1, \frac{1}{3} \cdot 2^{-k} - 1}\).
  Namely, \(\langle g_A(x), g_B(y) \rangle \leq \frac{1}{3} \cdot 2^{-k} - 1\).

- If \(f(x,y) = 1\), then \((g_A(x), g_B(y))\) is a YES instance of \(\text{GapInnerProduct}_{\frac{1}{3} \cdot 2^{-k} - 1, \frac{1}{3} \cdot 2^{-k} - 1}\).
  Namely, \(\langle g_A(x), g_B(y) \rangle \geq \frac{2}{3} \cdot 2^{-k} - 1\).

The following theorem gives an SMP protocol with \(\mu\)-randomness for GapInnerProduct (where \(\mu\) is any source with maximal correlation \(\rho\)). It matches the performance of the one-way protocol of [CGMS14].

Theorem 4.2.3 (SMP protocol for GapInnerProduct_{r,s}). Let \(\rho \in (0,1]\) and \(-1 \leq s < r \leq 1\) be given, and let \(\mu\) be any source of randomness with maximal correlation \(\rho\). There is an SMP protocol using \(\mu\)-randomness that solves \(\text{GapInnerProduct}_{r,s}^n\) using \(O\left(\frac{1}{\rho^2 (r-s)^2}\right)\) bits of communication.

We point out that Theorem 4.2.3 gives a protocol for sketching \(\ell_2\)-norms using imperfectly shared randomness, which might be of independent interest. Theorem 4.1.1 now follows by combining Lemma 4.2.2 and Theorem 4.2.3. In the rest of this section, we prove Theorem 4.2.3. First, we recall the following observation of [Wit75] which can be used to convert any source \(\mu\) of randomness with maximal correlation \(\rho\) to BGS(\(\rho\)).

Proposition 4.2.4 ([Wit75]). Let \(\mu\) be a source of randomness with maximal correlation \(\rho\). Given access to i.i.d. samples from \(\mu\), Alice and Bob can (without interaction) generate i.i.d. samples from BGS(\(\rho\)).

Proposition 4.2.4 follows from the definition of maximal correlation and from the two-dimensional Central Limit Theorem. We also recall the following well-known fact.

Fact 4.2.5 (Sheppard’s formula [She99]). If \((X, Y) \sim \text{BGS}(\rho)\) then \(\Pr[\text{Sign}(X) \neq \text{Sign}(Y)] = \frac{\arccos(\rho)}{\pi}\).

The following lemma is based on the well-known hyperplane rounding technique.
Lemma 4.2.6. Let \( \delta > 0 \) and \( \gamma < 0 \) be given, and let \( t = O(\log(1/\delta)/\gamma^2) \) be large enough. Let Alice be given \( (X_1, X_2, \ldots, X_t) \in \mathbb{R}^t \) and Bob be given \( (Y_1, Y_2, \ldots, Y_t) \in \mathbb{R}^t \) where \( (X_i, Y_i) \sim \text{BGS}(\eta) \) independently over \( i \in [t] \). Then, there is a deterministic SMP protocol that distinguishes the case where \( \eta \geq 0 \) from the case where \( \eta \leq \gamma \) using \( O(1/\gamma^2) \) bits of communication, and with probability at least \( 1 - \delta \) (where the probability is over \( (X_1, X_2, \ldots, X_t) \) and \( (Y_1, Y_2, \ldots, Y_t) \)).

Proof. For every \( i \in [t] \), Alice computes \( \tilde{X}_i = \text{Sign}(X_i) \) and sends the \( t \) bits \( \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_t \) to the referee. Similarly, Bob computes \( \tilde{Y}_i = \text{Sign}(Y_i) \) for each \( i \in [t] \), and sends the \( t \) bits \( \tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_t \) to the referee. Let \( \tau = (\arccos(\gamma)/\pi - 1/2)/2 \). The referee computes the Hamming distance \( \Delta(\tilde{X}, \tilde{Y}) \) and declares that \( \eta \geq 0 \) if \( \Delta(\tilde{X}, \tilde{Y}) \leq \tau \), and declares that \( \eta \leq \gamma \) otherwise. Note that if \( \eta \geq 0 \), then for each \( i \in [t] \),

\[
\Pr[\text{Sign}(X_i) \neq \text{Sign}(Y_i)] = \frac{\arccos(\eta)}{\pi} \leq \frac{\arccos(0)}{\pi} = \frac{1}{2}.
\]

On the other hand, if \( \eta \leq \gamma \), then for each \( i \in [t] \),

\[
\Pr[\text{Sign}(X_i) \neq \text{Sign}(Y_i)] = \frac{\arccos(\eta)}{\pi} \geq \frac{\arccos(\gamma)}{\pi} = \frac{1}{2} - \Theta(\gamma) - O(\gamma^3),
\]

where the last equality follows from the Taylor series approximation of \( \arccos(x) \) around \( x = 0 \). The proof now follows by combining Equations (4.1) and (4.2) and an application of the Chernoff bound.

We are now ready to prove Theorem 4.2.3.

Proof of Theorem 4.2.3. Alice is given \( u \in \mathbb{R}^n \) and Bob is given \( v \in \mathbb{R}^n \) such that \( \|u\|_2 = \|v\|_2 = 1 \). They are also given access to i.i.d. samples from a source \( \mu \) of randomness with maximal correlation \( \rho \). Using Proposition 4.2.4, Alice and Bob can (without any interaction) generate arbitrarily many i.i.d. samples from \( \text{BGS}(\rho) \). We first assume that \( r = 0 \). We will handle the more general case at the end of the proof. Set \( \gamma = \rho \cdot s \) and let \( t = O(\log(1/\delta)/\gamma^2) \) be as in the statement of Lemma 4.2.6. Draw \( t \) i.i.d vectors \( (X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(t)}, Y^{(t)}) \) each from \( \text{BGS}(\rho)^\otimes n \). Then, by elliptical symmetry, we get that independently over \( i \in [t], \langle (u, X^{(i)}), (v, Y^{(i)}) \rangle \sim \text{BGS}(\rho(\langle u, v \rangle)) \). Lemma 4.2.6 now implies an SMP protocol that distinguishes the case where \( \langle u, v \rangle \geq 0 \) from the case where \( \langle u, v \rangle \leq s \), using \( O(\frac{1}{\rho^2(r-s)^2}) \) bits of communication.

We now handle the case where \( r \) is not necessarily equal to 0. First, note that without loss of generality, we can assume that \( r \geq 0 \). This is because if \( r < 0 \), then Alice can negate each coordinate in her input vector which would preserve its \( \ell_2 \) norm and replace \( r \) by \( -s \geq 0 \) and \( s \) by \( -r \geq 0 \). Let \( N \triangleq n \cdot (1+r) \). Bob will construct a vector \( u' \in \mathbb{R}^N \), and Alice will construct a vector \( v' \in \mathbb{R}^N \), such that \( \|u'\|_2 = \|v'\|_2 = 1 \), and:

- If \( \langle u, v \rangle \geq r \), then \( \langle u', v' \rangle \geq 0 \).
- If \( \langle u, v \rangle \leq s \), then \( \langle u', v' \rangle \leq \frac{s+r}{1+r} = -\Theta(r-s) \).
To do so, Alice sets $u'_i = u_i \cdot \sqrt{n/N}$ for every $i \in [n]$ and $u'_i = +1/\sqrt{N}$ for all $i \in \{n + 1, \ldots, N\}$. On the other side, Bob sets $v'_i = v_i \cdot \sqrt{n/N}$ for all $i \in [n]$ and $v'_i = -1/\sqrt{N}$ for all $i \in \{n + 1, \ldots, N\}$. \qed
Chapter 5

Communication with Functional Uncertainty

5.1 Introduction & Related Work

In this chapter, we introduce our model for communication with functional uncertainty and present our corresponding protocols and limitation results.

Most forms of communication involve communicating players that share a large common context which they use to compress communication. In natural settings, the context may include understanding of the language and knowledge of the environment and laws. In designed (computer-to-computer) settings, the context includes “commonsense knowledge” as well as the knowledge of the operating system, communication protocols, and encoding/decoding mechanisms. This notion of “context” held by intelligent systems plays a fundamental role both in the classical study of artificial intelligence and in the emerging area of “conversational artificial intelligence” which underlies intelligent virtual assistants such as Siri, Google Assistant and Amazon Alexa. Remarkably, especially in the natural setting, context can seemingly be used to compress communication, even when it is enormous and not shared perfectly. This ability to communicate reliably despite a major source of uncertainty has led to a series of works attempting to model various forms of communication amid uncertainty, starting with those of Goldreich, Juba and Sudan [JS08], [GJS12] followed by [JKKS11], [JS11], [JW13], [HS14a] and [CGMS15]. The latter works implicitly give examples of context which share the three features mentioned above — the context helps compress communication, even though it is large and imperfectly shared. This chapter is the first work in this series to explicitly highlight this notion and features of context. It does so while studying a theme that is new to this series of works, namely a functional notion of uncertainty. We start by describing the setup for our model and then present our model and results below, before contrasting them with some of the previous works.

Our model builds on the classical setup of communication complexity due to Yao [Yao79]. The classical model considers two interacting players Alice and Bob each possessing some private information $X$ and $Y$, with $X$ known only to Alice and $Y$ only known to Bob.
In the general two-way setting, both players can send messages to each other, while in the one-way setting only Alice sends a message to Bob. They (specifically Bob, in the one-way setting) wish to compute some joint (Boolean-valued) function $g(X, Y)$ and would like to do so while communicating the minimum possible number of bits. In this chapter, we use the function $g$ to model (part of) the context of the communication. Indeed, it satisfies some of the essential characteristics of context. First, it is potentially "enormous". For example, if $g$ were represented as a truth table of values and if $X$ and $Y$ are $n$-bit strings, then the representation of $g$ would be $2^{2n}$ bits long. And indeed knowledge of this context can compress communication significantly: consider the trivial collection of examples where $g(X, Y) = g'(X)$, i.e., $g$ is simply a function of $X$. In this case, knowledge of the context (i.e., the function $g'$) compresses communication to just one bit. In contrast, if Alice does not know the context, her other option is to send $X$ to Bob which requires $n$ bits of communication. This intuitive explanation can be formalized using the well-known INDEXING problem [KN97, Example 4.19] which essentially considers the setting where Alice has an "index" (corresponding to $X$) and Bob has a vector (corresponding to the truth table of $g'$) and their goal is to compute the indexed value of the vector (i.e., computing $g'(X)$ in our correspondence). Standard lower bounds for INDEXING (see 5.10.2) imply that $\Omega(n)$ bits of communication are needed to compute $g'(X)$.

In our case, we focus on the case where the context is imperfectly shared. Specifically, we consider the setting where Bob knows the function $g$ and Alice only knows some (close) approximation $f$ to $g$ (with $f$ not being known to Bob). This leads to the questions: How should Alice and Bob interact while accounting for this uncertainty about their shared context? What quantitative effect does this uncertainty have on the communication complexity of computing $g(X, Y)$?

It is clear that if $X \in \{0, 1\}^n$, then $n$ bits of communication suffice — Alice can simply ignore $f$ and send $X$ to Bob. We wish to consider settings that improve on this. To do so, a necessary condition is that $g$ must have low communication complexity in the standard model. However, this necessary condition does not seem to be sufficient to compute $g$ correctly on every input — since Alice only has an approximation $f$ to $g$. (In 5.10.1 in Section 5.10, we formally prove this assertion by giving a function $g$ with low communication complexity, but where computing $g(X, Y)$ takes $\Omega(n)$ bits in the worst-case if Alice is only given an approximation $f$ to $g$.) Thus, we settle for a weaker goal, namely, that of computing $g$ correctly only on most inputs. This puts us in a distributional communication complexity setting. A necessary condition now is that $g$ must have a low-error low-communication protocol in the standard (distributional complexity) setting. The question is then: can $g$ be computed with low error and low communication when Alice only knows an approximation

---

1We note that the assumption that Bob knows the precise function $g$ to be computed is not a restrictive assumption but merely a convention that is consistent with our earlier suggestion that Bob wishes to compute the function $g$. We could have equally well asserted that the function to be computed is $f$ (and so only Alice knows the function to be computed), or picked a neutral setting saying that the function to be computed is $h$ which is very close to both $f$ and $g$. The definitions do not make a significant difference to the communication problem since any protocol $\Pi$ that computes a function close to $g$ is also close to $f$ or $h$, and hence all versions have the same communication complexity with small changes in the error probability.
f to g (with f being unknown to Bob)? Formalizing this model still requires some work and we do so next.

5.1.1 Uncertain-Communication Complexity

We first recall the standard model of communication complexity in the distributional setting. For contrast with our model, we sometimes refer to this as the model of “certain-communication”.

Let Π denote a communication protocol that specifies how Alice with input X and Bob with input Y interact, i.e., Π includes functions that specify: (1) given a history of transmissions, if the communication should continue and if so which one of Alice or Bob should speak next, (2) given a history of transmissions and the speaker’s private input (one of X or Y), what the speaker’s next message should be; and (3) what the output of the protocol is when the communication stops. We let Π(X, Y) denote the output of the protocol. Note that protocols may involve private or public (shared) randomness and if so Π(X, Y) is a random variable. We let the communication complexity of Π, denoted CC(Π), be the maximum number of bits exchanged by a protocol, maximized over all inputs and all (private or public) random coins. We say that a protocol is one-way if all communication comes from one speaker, typically from Alice to Bob.

In order to describe what it means for a protocol to compute a close approximation to a given function, we describe our distance measure on functions. For a distribution μ supported on \(\{0, 1\}^n \times \{0, 1\}^n\), we let \(\delta_\mu(f, g)\) denote the probability that \(f\) and \(g\) differ on a random input drawn from \(\mu\), i.e.,

\[
\delta_\mu(f, g) := \Pr_{(X, Y) \sim \mu}[f(X, Y) \neq g(X, Y)].
\]

If exactly one of \(f\) or \(g\) is probabilistic then we include the randomness in the probability space.\(^2\) We say \(f\) and \(g\) are \(\delta\)-close (with respect to \(\mu\)) if \(\delta_\mu(f, g) \leq \delta\).

For a parameter \(\epsilon > 0\), the distributional communication complexity (in the setting of certain-communication) of a function \(f : \{0, 1\}^n \times \{0, 1\}^n \to \{0, 1\}\) over a distribution \(\mu\), denoted \(\text{CC}_\mu^\epsilon(f)\), is the minimum communication complexity of a protocol, minimized over all protocols that compute a function that is \(\epsilon\)-close to \(f\), i.e.,

\[
\text{CC}_\mu^\epsilon(f) \triangleq \min_{\Pi : \delta_\mu(f, \Pi) \leq \epsilon} \{\text{CC}(\Pi)\}.
\]

Similarly, \(\text{owCC}_\mu^\epsilon(f)\) denotes the corresponding one-way communication complexity of \(f\).

We now turn to defining the measure of complexity in the uncertain setting. Ideally, we would like to define the uncertain-communication complexity of computing some function \(g\), given that Alice has some nearby function \(f\). But this definition will not make sense as such! Even if Alice does not know \(g\) the protocol might itself “know” \(g\). (Formally, the protocol \(\Pi\) that minimizes the communication complexity should not depend on \(g\), but how does one forbid this?) So the right formulation is to define the communication complexity of an entire

\(^2\)The correct generalization to the case when both \(f\) and \(g\) are probabilistic is to take the expectation of the statistical distance (also known as total variation distance) between \(f(X, Y)\) and \(g(X, Y)\), but we won’t need to consider this setting here.
family $\mathcal{F}$ of pairs of functions where $\mathcal{F} \subseteq \{(f, g) \mid f, g : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}\}$. We define such a measure shortly, but before doing so, we discuss one more aspect of uncertain-communication.

One view of communication, applicable in the uncertain-communication setting as well as the certain-communication setting, is to make the function being computed an explicit input to the communicating players, say by presenting it as a truth table. Thus, in the setting of uncertain-communication, we may view the goal as computing the universal function $U : ((f, X), (g, Y)) \mapsto g(X, Y)$, where Alice’s input is $(f, X)$ and Bob’s input is $(g, Y)$. In the certain-communication setting, we would further require $f = g$, but in our “uncertain” setting we do not. Instead, the functions $f, g$ are adversarially chosen subject to the restrictions that they are close to each other (under some distribution $\mu$ on the inputs) and that $g$ (and hence $f$) has a low-error low-communication protocol. The pair $(X, Y)$ is drawn from the distribution $\mu$ (independent of the choices of $f$ and $g$). The players both know $\mu$ in addition to their respective inputs.

Under this view, a protocol $\Pi$ solving an uncertain-communication problem is simply a protocol for the universal communication problem with its communication complexity being the maximum communication over all inputs and all possible randomness.\(^3\) The ability to solve communication problems from $\mathcal{F}$ under distribution $\mu$ is taken into account in defining the error of this protocol. For a protocol $\Pi$ computing a (probabilistic) function $\Pi((f, X), (g, Y))$, we let $\Pi_{(f,g)}$ denote the function $\Pi_{(f,g)}(X, Y) = \Pi((f, X), (g, Y))$.

**Definition 5.1.1 (Uncertain-Communication Complexity).** The (two-way) uncertain-communication complexity of a family $\mathcal{F}$ of pairs of functions $(f, g)$ with respect to a distribution $\mu$ supported on $\{0,1\}^n \times \{0,1\}^n$, denoted $\text{PubCC}_{\mu}(\mathcal{F})$, is the minimum communication cost of a public-coin protocol $\Pi$, such that for every $(f, g) \in \mathcal{F}$, the protocol $\Pi$ outputs $g(x, y)$ with probability at least $1 - \epsilon$ over the choice of $(x, y) \sim \mu$ and the public randomness. That is,

$$\text{PubCC}_{\mu}(\mathcal{F}) \triangleq \min_{\{\Pi \mid \forall (f,g) \in \mathcal{F} : \delta(\Pi(f,g), g) \leq \epsilon\}} \{\text{CC}(\Pi)\}.$$ 

We similarly define the (two-way) private-coin uncertain communication complexity $\text{PrivCC}_{\mu}(\mathcal{F})$ by restricting to private-coin protocols. The one-way measures $\text{owPubCC}_{\mu}(\mathcal{F})$ and $\text{owPrivCC}_{\mu}(\mathcal{F})$ are similarly defined by restricting to one-way protocols.

Note that we clearly have that $\text{PubCC}_{\mu}(\mathcal{F}) \leq \text{PrivCC}_{\mu}(\mathcal{F})$ and similarly for the one-way measures. The uncertain-communication complexity model is depicted in Figure 5-1.

**Remark 5.1.2.** The uncertain-communication model is clearly a generalization of Yao’s model which corresponds to the particular case where $\mathcal{F} = \{(f, f)\}$ for some fixed function $f$. On the other hand, the uncertain-communication model can also be viewed as a particular

\(^3\)It might seem more appropriate to define the communication as the maximum only over pairs $(f, g) \in \mathcal{F}$ and $(X, Y)$ in the support of $\mu$, but this does not make a difference for optimal protocols. A protocol can be modified to stop after a given number of bits of communication, and the result would only affect the accuracy of the output, which thereby becomes the only parameter tied to the problem being solved.
Our broad goal is to study $\text{PubCCU}_\epsilon(\mathcal{F})$ (and the corresponding private-coin and one-way measures) for a family $\mathcal{F}$, but this can be small only if the certain-communication complexity of functions in $\mathcal{F}$, specifically $\text{CC}_\epsilon(\mathcal{F}) \triangleq \max_{f,g \in \mathcal{F}} \{\text{CC}_\epsilon(g)\}$ (or the corresponding one-way measure), is small. Furthermore, we want to model “mild” uncertainty (and not total uncertainty) between Alice and Bob. To this end, we define the distance of a family $\mathcal{F}$, denoted by $\delta_\mu(\mathcal{F})$, to be the maximum over all $(f,g) \in \mathcal{F}$ of $\delta_\mu(f,g)$.

In what follows, we will study the behavior of $\text{PubCCU}_\epsilon(\mathcal{F})$ (and the corresponding private-coin and one-way measures) as a function of $\text{CC}_\epsilon(\mathcal{F})$ (or the corresponding one-way measure) and $\delta_\mu(\mathcal{F})$ and especially focus on the case where $\delta_\mu(\mathcal{F}) \ll \epsilon$ (so the uncertainty between Alice and Bob is very small compared to the error they are willing to tolerate).

### 5.1.2 Results

**Lower Bound on Public-Coin Protocols**  
For general distributions, it turns out we can prove a large gap between the public-coin uncertain-communication complexity of functions and their certain-communication complexity. Recall that for random variables $(X,Y)$ drawn from some joint distribution, the mutual information between $X$ and $Y$, denoted $I(X;Y)$, measures the amount of information that $X$ has about $Y$ (or vice versa).

**Theorem 5.1.3** (Lower Bound on Public-Coin Uncertain Protocols). For every constant $\delta \in (0,1)$ and $\epsilon \in (0,0.5)$, there exist constants $\tau > 0$ and $c < \infty$ such that for all positive integers $n$, there is a distribution $\mu$ supported on $\{0,1\}^n \times \{0,1\}^n$ and with mutual information $I \approx n$ along with a function class $\mathcal{F}$ satisfying $\delta_\mu(\mathcal{F}) \leq \delta$ and $\text{owCC}_0(\mathcal{F}) \leq 1$ such that $\text{PubCCU}_\epsilon(\mathcal{F}) \geq \tau \cdot \sqrt{I} - c$.

In particular, if $\delta$ is any positive constant (e.g., 0.001), then Theorem 5.1.3 asserts the existence of a distribution and a class of distance-$\delta$ functions for which the zero-error (one-way) communication complexity in the standard model is a single bit, but under contextual

---

4Formally, given a distribution $\mu$ over a pair $(X,Y)$ of random variables with marginals $\mu_X$ and $\mu_Y$ over $X$ and $Y$ respectively, the mutual information of $X$ and $Y$ is defined as $I(X;Y) \triangleq \mathbb{E}_{(a,b) \sim \mu} \left[ \log \left( \frac{\mu(a,b)}{\mu_X(a)\mu_Y(b)} \right) \right]$.

5We note that $I \approx n$ means that $I/n \to 1$ as $n \to \infty$. 

---

105
uncertainty, any two-way protocol (with an arbitrary number of rounds of interaction) having a noticeable advantage over random guessing requires $\Omega(\sqrt{n})$ bits of communication!

Theorem 5.1.3 above gives a lower bound of $\Omega(\sqrt{T})$ bits on the uncertain-communication, but this lower-bound does not grow with the uncertain communication $k$. The next theorem improves this lower bound to $\Omega(\sqrt{k} \cdot \sqrt{T})$ in the case of one-way communication.

**Theorem 5.1.4** (Improved Lower Bound on One-Way Public-Coin Uncertain Protocols). For every sufficiently small $\delta > 0$ and every positive integers $k, n$ such that $k = o(\exp(\sqrt{n}))$, there exist an input distribution $\mu$ on input pairs $(X,Y) \in \{0,1\}^{k\cdot n} \times \{0,1\}^{k\cdot n}$ with mutual information $I \approx k \cdot n$ and a function class $\mathcal{F} \triangleq \mathcal{F}_{\delta,k,n}$ such that:

(i) For each $(f,g) \in \mathcal{F}$, we have that $\delta(\mu,f,g) \leq \delta$.

(ii) For each $(f,g) \in \mathcal{F}$, we have that $\text{owCC}_\delta^\mu(f),\text{owCC}_\delta^\mu(g) \leq k$.

(iii) $\text{owPubCCU}_\delta^\mu(\mathcal{F}) = \Omega(\sqrt{k} \cdot \sqrt{T})$ for some absolute constant $\epsilon > 0$ independent of $\delta$.

As will be explained in detail in Section 5.7.1, the proof of Theorem 5.1.4 is based on an extension of the proof of the lower bound construction used in Theorem 5.1.3, which is then analyzed using additional novel techniques.

**Remark 5.1.5.** The construction that we use to prove Theorem 5.1.4 cannot give a lower bound larger than $\Theta(\sqrt{k} \cdot \sqrt{T})$. Thus, improving on the lower bound in Theorem 5.1.4 by more than logarithmic factors in $k$ and $I$ would require a new construction.

**One-Way Uncertain-Communication: Public-Coin Protocol** Given the strong negative result in Theorem 5.1.3 and Theorem 5.1.4, a natural question is to understand if there are any non-trivial settings where the uncertain-communication complexity is close to the certain-communication complexity. Surprisingly, it turns out that the uncertain-communication complexity can **always** be upper-bounded in terms of the certain-communication complexity and the mutual information of the input distribution. Theorem 5.1.6 shows that if $\mu$ is a distribution on which $f$ and $g$ are close and each has a one-way certain-communication complexity of at most $k$ bits (for all $(f,g) \in \mathcal{F}$), then the family $\mathcal{F}$ has one-way uncertain-communication complexity of at most $O(k \cdot (1+I))$ bits with $I$ being the mutual information of $(X,Y) \sim \mu$. We denote by $\text{CC}_\epsilon^\mu(\mathcal{F})$ (respectively, $\text{owCC}_\epsilon^\mu(\mathcal{F})$) the maximum over all $(f,g) \in \mathcal{F}$ of $\text{CC}_\epsilon^\mu(g)$ (respectively, $\text{owCC}_\epsilon^\mu(g)$).

We prove the following theorem.

**Theorem 5.1.6.** There exists a positive constant $c$ such that for all positive integers $k$ and $n$ and positive reals $\epsilon$, $\delta$ and $\theta$, for every distribution $\mu$ over $\{0,1\}^n \times \{0,1\}^n$, and every family $\mathcal{F}$ of pairs of Boolean functions satisfying $\delta(\mu,\mathcal{F}) \leq \delta$ and $\text{CC}_\epsilon^\mu(\mathcal{F}) \leq k$, it holds that

$$\text{owPubCCU}_{\epsilon+2\delta}(\mathcal{F}) \leq c \cdot \frac{(k + \log \frac{1}{\delta})}{\theta^2} \cdot \left(1 + \frac{I(X;Y)}{\theta^2}\right).$$

\[\text{(5.1)}\]

Note that if $\delta(\mu,f,g) \leq \delta$ and $\text{CC}_\epsilon^\mu(g) \leq k$, then $\text{CC}_{\epsilon+\delta}(f) \leq k$. The same statement holds similarly for the one-way measures.
Using the well-known fact that the one-way certain-communication of any function is at most exponential in its two-way communication complexity (e.g., [KN97, Exercise 4.21]), Theorem 5.1.6 also immediately implies the next corollary.

**Corollary 5.1.7.** There exists a positive constant $c$ such that for all positive integers $k$ and $n$ and positive reals $\epsilon$, $\delta$ and $\theta$, for every distribution $\mu$ over $\{0,1\}^n \times \{0,1\}^n$, and every family $\mathcal{F}$ of pairs of Boolean functions satisfying $\delta_\mu(\mathcal{F}) \leq \delta$ and $\text{CC}_\epsilon^\mu(\mathcal{F}) \leq k$, it holds that

$$
\text{owPubCCU}_{\epsilon+2\delta+\theta}(\mathcal{F}) \leq c \cdot \frac{(2^k + \log \left(\frac{1}{\theta}\right))}{\theta^2} \cdot \left(1 + \frac{I(X;Y)}{\theta^2}\right).
$$

(5.2)

We stress that the exponential blow-up in Equation (5.2) can be significantly smaller than the length $n$ of the inputs (which is the trivial upper bound on the communication in the uncertain-communication case). In the special case where $\mu$ is a product distribution, we have that $I(X;Y) = 0$, and we thus obtain the following particularly interesting corollary of Theorem 5.1.6.

**Corollary 5.1.8.** There exists a positive constant $c$ such that for all positive integers $k$ and $n$ and positive reals $\epsilon$, $\delta$ and $\theta$, for every product distribution $\mu$ over $\{0,1\}^n \times \{0,1\}^n$, and every family $\mathcal{F}$ of pairs of Boolean functions satisfying $\delta_\mu(\mathcal{F}) \leq \delta$ and $\text{owCC}_\epsilon^\mu(\mathcal{F}) \leq k$, it holds that

$$
\text{owPubCCU}_{\epsilon+2\delta+\theta}(\mathcal{F}) \leq c \cdot \frac{(k + \log \left(\frac{1}{\theta}\right))}{\theta^2}.
$$

In words, Corollary 5.1.8 says that for product distributions and for constant error probabilities, one-way uncertain-communication complexity is only a constant factor larger than the one-way certain-communication complexity.

One intuitive interpretation of the dependence on the mutual information $I(X;Y)$ in Theorem 5.1.6 is that the parties can make strong use of correlations among their inputs (i.e., between $X$ and $Y$) in the certain-communication setup. In contrast, they are unable to make such strong use in the uncertain case. Since the distribution $\mu$ in Theorem 5.1.3 has mutual information $\approx n$, Theorem 5.1.3 and Theorem 5.1.4 rule out improving the dependence on the mutual information in Theorem 5.1.6 to anything smaller than $\sqrt{I(X;Y)}$ in the two-way case, and to anything smaller than $\sqrt{k} \cdot \sqrt{I(X;Y)}$ in the one-way case. It is a very interesting open question to determine the correct exponent of $I(X;Y)$ and the right dependence on $k$ in Theorem 5.1.6.7

**One-Way Uncertain Communication: Imperfectly Shared Randomness Protocol**

The protocols underlying Theorem 5.1.6, Corollary 5.1.7 and Corollary 5.1.8 use *public randomness*, that is, they require Alice and Bob to perfectly share a long sequence of random

---

7We note that the upper bound of (roughly) $1+I(X;Y)$ on the communication blow-up due to uncertainty in Theorem 5.1.6 holds for *every* function class and input distribution whereas the lower bound of $\sqrt{I}$ given in Theorem 5.1.3 holds for *some* function class and input distribution (and similarly for the lower bound given in Theorem 5.1.4). In particular, if the distribution $\mu$ only puts mass on points $(X,Y)$ for which $X = Y$, then the mutual information can be very large while there would be no blow-up in communication due to uncertainty (since on such distributions no communication is needed to compute any function).
bits (which, intuitively, they use in order to “fool” the adversary who selects the pair \((f, g)\) of functions). But can Alice and Bob achieve similarly efficient communication in the case where they have access to \textit{weaker forms of correlation}?

We point out that understanding the type of randomness that is needed in order to cope with uncertainty is a core question in the setup of communication with contextual uncertainty: \textit{If Alice and Bob do not (perfectly) agree on the function being computed, why can they be assumed to (perfectly) agree on the shared randomness?}

It turns out that in the case of product distributions, i.e., in the setting of Corollary 5.1.8, it is not necessary for Alice and Bob to \textit{perfectly} share the sequence of random bits. If Alice is given a uniform-random string \(r\) of bits and Bob is given a string \(r'\) obtained by independently flipping each coordinate of \(r\) with probability \(\frac{1}{2}\), then efficient communication is still possible!

More formally, for \(\rho \in [0, 1]\), define \(\text{owIsrCCU}_{\epsilon, \rho}(\mathcal{F})\) in the same way that we defined \(\text{owPubCCU}_{\epsilon}(\mathcal{F})\) except that instead of Alice and Bob having access to public randomness, Alice will have access to a sequence \(r\) of independent uniformly-random bits, and Bob will have access to a sequence \(r'\) of bits obtained by independently flipping each coordinate of \(r\) with probability \((1 - \rho)/2\). Note that this setup of imperfectly shared randomness “interpolates between” the public randomness and private randomness setups, i.e., \(\text{owIsrCCU}_{\epsilon, 1}(\mathcal{F}) = \text{owPubCCU}_{\epsilon}(\mathcal{F})\) and \(\text{owIsrCCU}_{\epsilon, 0}(\mathcal{F}) = \text{owPrivCCU}_{\epsilon}(\mathcal{F})\).

\textbf{Theorem 5.1.9} (Uncertain Protocol using Imperfectly Shared Randomness). Let \(\rho \in (0, 1]\) and \(\mu\) be a product distribution. Let \(\mathcal{F}\) consist of pairs \((f, g)\) of functions with \(\delta_\mu(f, g) \leq \delta\), and \(\text{owCC}_\epsilon^\mu(f), \text{owCC}_\epsilon^\mu(g) \leq k\). Then, for every positive \(\theta\), \(\text{owIsrCCU}_{\epsilon + 2k + \theta, \rho}(\mathcal{F}) \leq O_\theta(k/\rho^2)\).

The \textit{imperfectly shared randomness} model in Theorem 5.1.9 was recently independently introduced (in the setup of communication complexity) by Bavarian, Gavinsky and Ito [BGI14] and by Canonne, Guruswami, Meka and Sudan [CGMS15] (and it was further studied in [GKS16a]). Moreover, our proof of Theorem 5.1.9 is based on combining the ideas behind the uncertain-communication protocol in Corollary 5.1.8 and the locality sensitive hashing based protocol of [CGMS15].

We point out that Theorem 5.1.9 also holds for more general i.i.d. sources of correlated randomness than the one described above. More precisely, for i.i.d. (not necessarily binary) sources of (imperfectly) shared randomness with \textit{maximal correlation} \(\rho\), the work of Wit- senhausen [Wit75] along with the same ideas behind the protocol in Theorem 5.1.9 imply an uncertain-communication protocol with \(O_\theta(k/\rho^2)\) bits of communication.

\textbf{Hardness of Contextual Agreement} We point out that our results in Theorem 5.1.6, Corollary 5.1.7 and Corollary 5.1.8 achieve reliable communication despite uncertainty about the context \textit{even when the uncertainty itself is hard to resolve}. To elaborate on this statement,

---

\[^{8}\text{The \textit{maximal correlation} of a pair \((X, Y)\) of random variables (with support \(\mathcal{X} \times \mathcal{Y}\)) is defined as} \rho(X, Y) \triangleq \sup \mathbb{E}[F(X)G(Y)] \text{ where the supremum is over all functions } F : \mathcal{X} \to \mathbb{R} \text{ and } G : \mathcal{Y} \to \mathbb{R} \text{ with } \mathbb{E}[F(X)] = \mathbb{E}[G(Y)] = 0 \text{ and } \text{Var}[F(X)] = \text{Var}[G(Y)] = 1. \text{ It is not hard to show that the binary source of imperfectly shared randomness defined before Theorem 5.1.9 has maximal correlation } \rho.\]
note that one hope for achieving a low-communication protocol for \( g \) would be for Alice and Bob to first agree on some function \( h \) that is close to \( f \) and \( g \), and then apply some low-communication protocol for this common function \( h \). Such a protocol obviously exists if we assume that \( g \) has a low-communication protocol, albeit with slightly higher error. (In particular, an \( \epsilon \)-error protocol for \( g \) computes \( h \) with error \( \epsilon + \delta_{\mu}(g, h) \).) This would be the “resolve the uncertainty first” approach.

We prove (in Theorem 5.1.10 below) that resolving the uncertainty is definitely an overkill and can lead to communication exponential in \( n \) (and much more so than the trivial protocol of sending \( x \)) and hence, this cannot be a way to prove Theorem 5.1.6. Namely, denote by \( \text{AGREE}_{\delta, \gamma}(\mathcal{F}) \) the communication problem where Alice gets \( f \) and Bob gets \( g \) such that \((f, g) \in \mathcal{F}\) and their goal is for Alice to output \( h_A \) and Bob to output \( h_B \) such that \( \delta(h_A, f), \delta(h_B, g) \leq \delta \) and \( \Pr[h_A = h_B] \geq \gamma \), where the probability is over the internal randomness of the protocol. Even getting a positive agreement probability \( \gamma \), let alone getting agreement with high probability, turns out to require high communication as shown by the following theorem.

**Theorem 5.1.10.** Let \( \mu \) denote the uniform distribution over \( \{0, 1\}^n \times \{0, 1\}^n \). For every \( \delta, \delta' \in (0, 1/2) \) and \( \gamma \in (0, 1) \), there exist \( \alpha > 0 \) and \( \beta < \infty \) and a family \( \mathcal{F} \) of pairs of Boolean functions satisfying \( \delta_{\mu}(\mathcal{F}) \leq \delta \) and \( \text{CC}_{\mu}(\mathcal{F}) = 0 \), such that

\[
\text{CC}(\text{AGREE}_{\delta', \gamma}(\mathcal{F})) \geq \alpha \cdot 2^n - \beta.
\]

In particular, the theorem shows that there is a class of function pairs \((f, g)\) where \( f \) and \( g \) are very close (say \( \delta(f, g) \leq 0.01 \)) but agreeing on a function \( h \) with even a slight correlation with \( f \) and \( g \) (say \( \delta(f, h), \delta(g, h) \leq 0.499 \)) incurs an exponentially high communication cost in \( n \). We note that Theorem 5.1.10 holds in the case where Alice and Bob have access to an unlimited amount of shared randomness.

**Lower Bound on Private-Coin Protocols** In light of the uncertain-communication protocol with imperfectly shared randomness given in Theorem 5.1.9, it is natural to ask whether it is possible to obtain a blow-up which is at most a constant factor via a private-coin protocol. Note that this corresponds to setting \( \rho = 0 \) in Theorem 5.1.9, in which case that theorem provides no useful bound.

Our next result (Theorem 5.1.11) shows that private-coin protocols are much weaker than imperfectly shared randomness protocols (and hence much weaker than public-coin protocols) in the setup of communication with contextual uncertainty. Far from obtaining a constant factor blow-up in communication, private-coin protocols incur an increase that is a growing function of \( n \) when dealing with uncertainty.

Let \( \mathcal{U} \triangleq \mathcal{U}_2^n \) be the uniform distribution on \( \{0, 1\}^{2n} \). For positive integers \( t \) and \( n \), we define \( \log^{(i)}(n) \) by setting \( \log^{(1)}(n) = \log n \), and \( \log^{(i)}(n) = \max(\log \log^{(i-1)}(n), 1) \) for all \( i \in \{2, \ldots, t\} \).

**Theorem 5.1.11 (Lower Bound on Private-Coin Uncertain Protocols).** For every sufficiently small \( \delta > 0 \), there exist a positive integer \( \ell \triangleq \ell(\delta) \) and a function class \( \mathcal{F} \triangleq \mathcal{F}_\delta \) such that:
(i) For each \((f, g) \in \mathcal{F}\), we have that \(\delta_{\mathcal{U}}(f, g) \leq \delta\).

(ii) For each \((f, g) \in \mathcal{F}\), we have that \(\text{owCC}_0^\ell(f), \text{owCC}_0^\ell(g) \leq \ell\).

(iii) For every \(\eta > 0\) and \(\epsilon \in (4\delta, 0.5]\), we have that \(\text{PrivCCU}_{\epsilon/2-2\delta-\eta}(\mathcal{F}) = \Omega(\eta^2 \cdot \log(t) \cdot (n))\) for some positive integer \(t = \Theta((\epsilon/\delta)^2)\).

We note that Theorem 5.1.10 and Theorem 5.1.11 together imply that it is necessary for Alice and Bob to have access to some form of correlation in order to incur no more than a constant factor blow-up in communication for product distributions.

In Theorem 5.1.11, the inputs \(x\) and \(y\) are binary strings of length \(n\) and \(\mathcal{F}\) is a family of pairs of functions, which each function mapping \(\{0, 1\}^n \times \{0, 1\}^n\) to \(\{0, 1\}\). Also, the parameter \(\eta\) can possibly depend on \(n\). We point out that Theorem 5.1.11 also implies the first separation between deterministic uncertain protocols and public-coin uncertain protocols\(^9\).

Remark 5.1.12. We point out that the relative power of private-coin and public-coin protocols in the uncertain model is both conceptually and technically different from the standard model. Specifically, the randomness is potentially used in the standard model in order to fool an adversary selecting the input pair \((x, y)\), whereas in the uncertain model, it is used to fool an adversary selecting the pair \((f, g)\) of functions that are promised to be close. This promise makes the task of proving lower bounds against private-coin protocols in the uncertain model (e.g., Theorem 5.1.11) significantly more challenging than in the standard model.\(^{10}\) Moreover, a well-known theorem due to Newman [New91] shows that in the standard model, any public-coin protocol can be simulated by a private-coin protocol while increasing the communication by an additive \(O(\log n)\) bits. By contrast, there is no known analogue of Newman’s theorem in the uncertain case!

Remark 5.1.13. The construction that we use to prove Theorem 5.1.11 cannot give a separation larger than \(\Theta(\log \log n)\). Thus, showing a separation of \(\omega(\log \log n)\) between private-coin and public-coin protocols in the uncertain case would require a new construction. For more details, see Remark 5.8.13.

We next discuss some conceptual implications of our results.

Remark 5.1.14. As mentioned in Remark 5.1.2, the uncertain model is clearly a generalization of Yao’s model. Strictly speaking, the uncertain model can also be viewed as a particular case of Yao’s model by regarding the function(s) that is being computed as part of the inputs of Alice and Bob, which results in an exponential blow-up in the input-size. This latter view turns out to be fruitless for our purposes. Indeed, from this perspective, all the different well-studied communication functions (such as Equality, Set Disjointness, Pointer

\(^9\)This uses the fact that private-coin communication complexity is no larger than deterministic communication complexity, both in the certain and uncertain setups.

\(^{10}\)In particular, the diagonalization-based arguments that imply a separation between the public-coin and the private-coin communication complexities of the Equality function in the standard model completely fail when we impose such a promise.
Jumping, etc.) are regarded as special cases of one “universal function”! More importantly, this view completely blurs the distinction between the goal of the communication (i.e., the function to compute) and the inputs of the parties. On a technical level, it does not simplify the task of proving the lower bounds in Theorems 5.1.10, 5.1.11 and 5.1.4 in any way since it does not capture the promise that the two functions (given to Alice and Bob) are close in Hamming distance. Thus, we henceforth stick to the former view and use the expressions “uncertain model” and “standard model” to refer to the setups with and without uncertainty, respectively.

5.1.3 Prior Work

The first works to consider uncertain goal-oriented communication in a manner similar to ours were those of [JS08] and [GJS12]. Their aim was to model an extreme form of uncertainty, where Alice and Bob do not have any prior (known) commonality in context and indeed both come with their own “protocol” which tells them how to communicate. So communication is needed even to resolve this uncertainty. While their setting is thus very broad, the solutions they propose are less communication-efficient and typically involve resolving the uncertainty as a first step.

The later works [JKKS11], [HS14a] and [CGMS15] tried to restrict the forms of uncertainty to see when it could lead to more efficient communication solutions. For instance, [JKKS11] consider the compression problem when Alice and Bob do not completely agree on the prior. This introduces some uncertainty in the beliefs, and they provide fairly efficient solutions by restricting the uncertainty to a manageable form (this question of uncertain compression is studied in detail in a distributed setup in Chapter 6 of this thesis). [CGMS15] were the first to connect this stream of work to communication complexity which seems to be a good umbrella to study the broader uncertain communication problems. The imperfectness they study is however restricted to the randomness shared by the communicating parties, and does not incorporate any other elements. (We point out that the setup of communication with imperfectly shared randomness had independently been studied by [BGI14] in the simultaneous message passing model. It was also further studied by [GKS16a]). [CGMS15] suggest studying imperfect understanding of the function being computed as a general direction, though they do not suggest any specific definitions, which we in particular do in this chapter.

Organization In Section 5.2, we carefully develop the uncertain communication complexity model after recalling the standard distributional communication complexity model. In Section 5.3, we prove the hardness of contextual agreement (Theorem 5.1.10). In Section 5.4, we prove our upper bound for public-coin protocols (Theorem 5.1.6), and in Section 5.5, we prove our upper bound for imperfectly shared randomness protocols (Theorem 5.1.9). In Section 5.6, we prove our lower bound for two-way public-coin protocols (Theorem 5.1.3), and in Section 5.7, we proved our improved lower bound for one-way public-coin protocols (Theorem 5.1.4). In Section 5.8, we prove our lower bound on private-coin protocols (The-
orem 5.1.11). A useful lemma that is used in Section 5.8 appears in Appendix 5.9. In Section 5.11, we conclude with some interesting open questions.

5.2 The Uncertain-Communication Complexity Model

We start with some general notation. For a positive integer \( n \), we let \( [n] \triangleq \{1, \ldots, n\} \). For a real number \( x \), we define \( \text{Sign}(x) \) to be 1 if \( x \geq 0 \) and 0 if \( x < 0 \). For a real number \( x \), we also denote \( \exp(x) \) a quantity of the form \( 2^{\Theta(x)} \), and we use \( \log x \) to denote a logarithm in base 2. For a distribution \( \mu \), we denote by \( X \sim \mu \) the process of sampling a random variable from the distribution \( \mu \). For a set \( S \), we write \( X \in R_S \) to indicate that \( X \) is a random variable that is uniformly distributed on \( S \). For any two subsets \( S, T \subseteq [n] \), we let \( S \setminus T \) be the set of all elements of \( S \) that are not in \( T \). We let \( S \triangle T \) be the symmetric difference of \( S \) and \( T \), i.e., the union of \( S \setminus T \) and \( T \setminus S \). We next recall the classical communication complexity model of [Yao79] and then present our definitions and measures.

5.2.1 Communication Complexity

We give the basic definitions related to communication complexity. A more extensive treatment can be found in [KN97]. Let \( f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \) be a function and Alice and Bob be two parties. A protocol \( \Pi \) between Alice and Bob specifies how and what Alice and Bob communicate given their respective inputs and communication thus far. It also specifies when they stop and produce an output (that we require to be produced by Bob). A protocol is said to be one-way if it involves a single message from Alice to Bob, followed by Bob producing the output. The protocol \( \Pi \) is said to compute the function \( f \) if for every \( (x, y) \in \mathcal{X} \times \mathcal{Y} \) it holds that \( \Pi(x, y) = f(x, y) \). The communication complexity of \( f \) is the minimal communication cost of a protocol computing \( f \).

It is usual to relax the above setting by introducing a distribution \( \mu \) over the input space \( \mathcal{X} \times \mathcal{Y} \) and requiring the protocol to succeed with high probability (rather than with probability 1). Specifically, we say that a protocol \( \Pi \) \( \epsilon \)-computes a function \( f \) under a
distribution μ if δμ(Π, f) ≤ ε. We next define the distributional communication complexity both for functions (as usual in the field of communication complexity) and for families of pairs of functions (which, as discussed in Section 5.1, are central to our work in this chapter).

**Definition 5.2.1** (Distributional Communication Complexity). Let \( f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \) be a Boolean function and μ be a probability distribution over \( \mathcal{X} \times \mathcal{Y} \). The distributional communication complexity of \( f \) under μ with error ε, denoted by \( \text{CC}_\mu(\epsilon)(f) \), is defined as the minimum over all protocols Π that \( \epsilon \)-compute \( f \) over μ, of the communication cost of Π. The one-way distributional communication complexity \( \text{owCC}_\mu(\epsilon)(f) \) is defined similarly by minimizing over one-way protocols Π.

Let \( \mathcal{F} \subseteq \{ f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}\}^2 \) be a family of pairs of Boolean functions with domain \( \mathcal{X} \times \mathcal{Y} \). We define the distributional communication complexity \( \text{CC}_\mu(\epsilon)(\mathcal{F}) \) of \( \mathcal{F} \) as the maximum value of \( \text{CC}_\mu(\epsilon)(g) \) over all pairs \( (f, g) \in \mathcal{F} \). Similarly, we define the one-way distributional communication complexity \( \text{owCC}_\mu(\epsilon)(\mathcal{F}) \) of \( \mathcal{F} \) as the maximum value of \( \text{owCC}_\mu(\epsilon)(g) \) over all functions \( (f, g) \in \mathcal{F} \).

We note that it is also common to provide Alice and Bob with a shared random string which is independent of \( x, y \) and \( f \). In the distributional communication complexity model, it is a known fact [Yao77] that any protocol with shared randomness can be used to get a protocol that does not use shared randomness without increasing its distributed communication complexity.

We next give the standard definition of (worst-case) communication complexity.

**Definition 5.2.2** (Deterministic Communication Complexity). The two-way (respectively, one-way) deterministic communication complexity of \( f \), denoted by \( \text{CC}(f) \) (respectively, \( \text{owCC}(f) \)), is defined as the minimum over all two-way (respectively, one-way) deterministic protocols Π that compute \( f \) correctly on every input pair, of the communication cost of Π.

**Definition 5.2.3** (Private-Coin Communication Complexity). The two-way (respectively, one-way) private-coin communication complexity of \( f \) with error ε, denoted by \( \text{PrivCC}_\epsilon(\epsilon)(f) \) (respectively, \( \text{owPrivCC}_\epsilon(\epsilon)(f) \)), is defined as the minimum over all two-way (respectively, one-way) private-coin protocols Π that compute \( f \) with probability at least \( 1 - \epsilon \) on every input pair, of the communication cost of Π.

The measures given in Definitions 5.2.2 and 5.2.3 can be similarly defined for partial functions \( f \).

### 5.2.2 Uncertain-Communication Complexity

We now turn to the central definition of this chapter: uncertain-communication complexity. Our goal is to understand how Alice and Bob can communicate when the function that Bob wishes to determine is not known to Alice. In this setting, we make the functions \( g \) (that Bob wants to compute) and \( f \) (Alice’s estimate of \( g \)) explicitly part of the input to the protocol.
Π. Thus, in this setting, a protocol Π specifies how Alice with input (f, x) and Bob with input (g, y) communicate, and how they stop and produce an output. We denote the output by Π((f, x), (g, y)). We say that Π computes (f, g) if for every (x, y) ∈ 𝒳 × 𝒴, the protocol outputs g(x, y). We say that a (possibly public-coin) protocol Π ϵ-computes (f, g) over μ if Pr[g(X, Y) ≠ Π((f, X), (g, Y))] ≤ ϵ where the probability is over (X, Y) ∼ μ (and possibly over the public randomness of the protocol).

Next, one may be tempted to define the communication complexity of a pair of functions (f, g) as the minimum over all protocols that compute (f, g) of their maximum communication. But this does not capture the uncertainty! (Rather, a protocol that works for the pair corresponds to both Alice and Bob knowing both f and g.) To model the uncertainty, we have to consider the communication complexity of a whole class of pairs of functions, from which the pair (f, g) is chosen (in our case by an adversary).

Let ℱ ⊆ {f : 𝒳 × 𝒴 → {0, 1}}² be a family of pairs of Boolean functions with domain 𝒳 × 𝒴. We say that a public-coin protocol Π ϵ-computes ℱ over μ if for every (f, g) ∈ ℱ, we have that Π ϵ-computes (f, g) over μ.

We now define the uncertain-communication complexity of a family ℱ of functions. (Note that this is exactly the same as Definition 5.1.1.)

**Definition 5.2.4** (Uncertain-Communication Complexity). Let μ be a distribution on 𝒳 × 𝒴 and ℱ ⊆ {f : 𝒳 × 𝒴 → {0, 1}}². The (two-way) uncertain-communication complexity of ℱ with respect to μ, denoted PubCCUμ(ℱ), is the minimum communication cost of a public-coin protocol Π, such that for every (f, g) ∈ ℱ, the protocol Π outputs g(x, y) with probability at least 1 − ϵ over the choice of (x, y) ∼ μ and the public randomness. That is,

\[
\text{PubCCU}_μ(ℱ) \triangleq \min_{\{\Pi \mid \forall (f, g) \in ℱ, \delta(\Pi(f, g), g) \leq \epsilon\}} \{\text{CC}(\Pi)\}.
\]

We similarly define the (two-way) private-coin uncertain communication complexity PrivCCUμ(ℱ) by restricting to private-coin protocols. The one-way measures owPubCCUμ(ℱ) and owPrivCCUμ(ℱ) are similarly defined by restricting to one-way protocols.

We remark that while in the standard distributional model of Section 5.2.1, the “easy direction” of Yao’s minimax principle [Yao77] implies that shared randomness can be assumed without loss of generality, this is not necessarily the case in Definition 5.2.4. This is because the function pair (f, g) is selected adversarially from the class ℱ and hence shared randomness can help the protocol “fool” this adversary.¹¹

Also, observe that in the special case where ℱ = {(f, g)} consists of a single pair of functions, Definition 5.2.4 reduces to the standard definition of distributional communication complexity (i.e., Definition 5.2.1) for the function-class ℱ = {(f, g)}, and we thus have

¹¹If the pair (f, g) was sampled from some fixed probability distribution, then shared randomness would no longer be needed and deterministic protocols would be optimal. However, an adversarial assumption on (f, g) is more desirable since it is more likely to model natural scenarios. The reason why we choose the input pair (x, y) from a fixed distribution is to be able to define a notion of distance between two functions. Henceforth, we assume that the pair (f, g) is chosen adversarially.
PubCCUμ(\{(f, g)\}) = CCμ(\{(f, g)\}). Furthermore, the uncertain communication complexity is monotone, i.e., if \(\mathcal{F} \subseteq \mathcal{F}'\) then PubCCUμ(\mathcal{F}) \leq PubCCUμ(\mathcal{F}'). Hence, we conclude that PubCCUμ(\mathcal{F}) \geq CCμ(\mathcal{F}).\footnote{The analogous inequalities also hold for private-coin and one-way protocols.}

Furthermore, note that we have to allow uncertain protocols to have positive error probabilities as shown in Section 5.10.

In this chapter, we attempt to identify a setting under which the last lower bound above can be matched. If the set of functions \(\Gamma(g) := \{f \mid (f, g) \in \mathcal{F}\}\) is not sufficiently informative about \(g\), then it seems hard to conceive of settings where Alice and Bob can do non-trivially well. We thus impose a simple and natural restriction on \(\Gamma(g)\), namely, that it consists of functions that are close to \(g\) (in \(\delta_\mu\)-distance). This leads us to the definition of the distance of a family of pairs of functions.

**Definition 5.2.5** (Distance of a family, \(\delta_\mu(\mathcal{F})\)). Let \(\mathcal{F} \subseteq \{f : \mathcal{X} \times \mathcal{Y} \rightarrow \{0, 1\}\}^2\) be a family of pairs of Boolean functions with domain \(\mathcal{X} \times \mathcal{Y}\), and let \(\mu\) be a distribution over \(\mathcal{X} \times \mathcal{Y}\). The \(\mu\)-distance of \(\mathcal{F}\), denoted \(\delta_\mu(\mathcal{F})\), is defined as the maximum over all \((f, g) \in \mathcal{F}\) of the distance \(\delta_\mu(f, g)\).

An optimistic hope might be that given \((f, g)\) the players can exchange a few bits and agree on a function \(h\) which is close to both \(f\) and \(g\), and thus reduce the task to that of computing \(h\) in the standard (certain-communication) setting. Our Theorem 5.1.10 shows that this naive strategy cannot work, in that there exists a family of nearby functions where agreement takes exponentially more communication than the simple strategy of simply exchanging \(x\) and \(y\). We then prove Theorem 5.1.6 which in particular gives an upper bound on the one-way uncertain-communication complexity, \(\mathsf{owPubCCU}_\mu(\mathcal{F})\), which is comparable\footnote{up to a small increase in the error probability} to the one-way certain-communication complexity \(\mathsf{owCC}_\mu(\mathcal{F})\), when \(\delta_\mu(\mathcal{F})\) is small, and \(\mu\) is a product distribution (Corollary 5.1.8). We also show that a similar upper bound holds for weaker protocols, namely, those only having access of imperfectly shared randomness (Theorem 5.1.9). More generally, Theorem 5.1.6 shows that, in the case of public-coin protocols, the bound grows slowly as long as the mutual information between \(X\) and \(Y\) is small. We also prove Theorem 5.1.3 and Theorem 5.1.4, showing that for general non-product distributions, \(\mathsf{PubCCU}_\mu(\mathcal{F})\) and \(\mathsf{owPubCCU}_\mu(\mathcal{F})\) can be much larger than \(\mathsf{owCC}_\mu(\mathcal{F})\) even when the distance \(\delta_\mu(\mathcal{F})\) is a small constant. For instance, in Theorem 5.1.3, we construct a family of nearby functions along with a distribution \(\mu\) for which the one-way certain-communication complexity is a single bit whereas the two-way uncertain-communication complexity is at least \(\Omega(\sqrt{n})\) bits. We also prove Theorem 5.1.11 which shows that \(\mathsf{PrivCCU}_\mu(\mathcal{F})\) can be much larger than both \(\mathsf{owCC}_0(\mathcal{F})\) and \(\mathsf{owPubCCU}_\mu(\mathcal{F})\).

### 5.3 Hardness of Contextual Agreement

In this section, we show that even if \(f\) and \(g\) are very close and have small one-way distributional communication complexity over a distribution \(\mu\) (for every \((f, g) \in \mathcal{F}\), agreeing
on an $h$ such that $\delta_\mu(h, f)$ and $\delta_\mu(g, h)$ are non-trivially small takes communication that is roughly the size of the binary representation of $f$ (which is exponential in the size of the input). Thus, agreeing on $h$ before simulating a protocol for $h$ is exponentially costlier than even the trivial protocol where Alice sends her input $x$ to Bob. Formally, we consider the following communication problem:

**Definition 5.3.1** ($\text{AGREE}_{\delta, \gamma}(\mathcal{F})$). For any given family of pairs of functions $\mathcal{F} \subseteq \{ f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}\}^2$, the $\mathcal{F}$-agreement problem with parameters $\delta, \gamma \geq 0$ (denoted by $\text{AGREE}_{\delta, \gamma}(\mathcal{F})$) is the communication problem where Alice gets $f$ and Bob gets $g$ such that $(f, g) \in \mathcal{F}$ and their goal is for Alice to output $h_A$ and Bob to output $h_B$ such that $\delta(h_A, f), \delta(h_B, g) \leq \delta$ and $\Pr[h_A = h_B] \geq \gamma$, where the probability is over the internal randomness of the protocol.

The contextual agreement problem is illustrated in Figure 5-2 (where $C$ denotes the number of bits of communication).

![Figure 5-2: Contextual Agreement Problem](image)

Somewhat abusing notation, we will use $\text{AGREE}_{\delta, \gamma}(\mathcal{D})$ to denote the distributional problem where $\mathcal{D}$ is a distribution on $\{ f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}\}^2$ and the goal now is to get agreement with probability $\gamma$ over the randomness of the protocol and that of the inputs.

If the agreement problem could be solved with low communication for a family $\mathcal{F}$ of pairs of Boolean functions, then it would imply a natural protocol for $\mathcal{F}$ in the uncertain-communication case. The following theorem, which is a refinement of Theorem 5.1.10 proves that agreement is extremely expensive even when all the functions that appear in the class $\mathcal{F}$ have zero communication complexity.

**Theorem 5.3.2.** Let $\mu$ denote the uniform distribution over $\mathcal{X} \times \mathcal{Y}$. For every $\delta, \delta' \in (0, 1/2)$, there exist $\alpha > 0$ and $\beta < \infty$ such that for every $\gamma > 0$ and finite sets $\mathcal{X}$ and $\mathcal{Y}$, the following holds: There is a family $\mathcal{F}$ of pairs of Boolean functions over $\mathcal{X} \times \mathcal{Y}$ satisfying $\delta_\mu(\mathcal{F}) \leq \delta$ and $\text{CC}^0(\mathcal{F}) = 0$, such that $\text{CC}(\text{AGREE}_{\delta', \gamma}(\mathcal{F})) \geq \alpha|\mathcal{Y}| - \beta \log(1/\gamma)$ where $\mu$ is the uniform distribution over $\mathcal{X} \times \mathcal{Y}$.

Note that Theorem 5.1.10 corresponds to the special case of $\mathcal{X} = \mathcal{Y} = \{0, 1\}^n$ and $\gamma$ being an absolute constant.

Theorem 5.3.2 says that there is a family of pairs of functions supported on functions of zero communication complexity (with zero error) for which agreement takes communication
polynomial in the size of the domain of the functions. Note that this is exponentially larger than the trivial communication complexity upper bound for any function \( g \), which is at most \( \min\{1 + \log |\mathcal{Y}|, \log |\mathcal{X}|\} \) (and which would result from either Alice sending the binary representation of her input to Bob, or Bob sending the binary representation of his input to Alice). Furthermore, this lower bound holds even if the goal is to get agreement with probability only exponentially small in \(|\mathcal{Y}|\), which is really tiny!

Our proof of Theorem 5.3.2 uses a lower bound on the communication complexity of the agreement distillation (with imperfectly shared randomness) problem defined in \cite{CGMS15}, which in turn relies on the lower bound of \cite{BM11} on common randomness generation from correlated sources in the zero-communication case. We describe the problem of \cite{CGMS15} below and the result that we use. We note that their context is slightly different and our description below is a reformulation. First, we define the notion of \( \eta \)-noisy sequences of bits.

A pair of bits \((a, b)\) is said to be a pair of \( \eta \)-noisy uniform bits if \( a \) is uniform over \( \{0, 1\} \), and \( b = a \) with probability \( 1 - \eta \) and \( b \neq a \) with probability \( \eta \). A pair of sequences of bits \((r, s)\) is said to be \( \eta \)-noisy if \( r = (r_1, \ldots, r_n) \) and \( s = (s_1, \ldots, s_n) \) and each coordinate-pair \((r_i, s_i)\) is an \( \eta \)-noisy uniform pair drawn independently of all other pairs. For a random variable \( W \), we define its min-entropy as \( H_\infty(w) \triangleq \min_{w \in \text{supp}(W)}\{-\log(\Pr[W = w])\} \).

**Definition 5.3.3 (Agreement-Distillation\(^k_{\gamma, \eta}\)).** In this problem, Alice and Bob get as inputs \( r \) and \( s \) respectively, where \((r, s)\) is an \( \eta \)-noisy sequence of bits. Their goal is to communicate deterministically and produce as outputs \( w_A \) (Alice’s output) and \( w_B \) (Bob’s output) with the following properties: (i) \( H_\infty(w_A), H_\infty(w_B) \geq k \) and (ii) \( \Pr_{(r, s)}[w_A = w_B] \geq \gamma \).

**Lemma 5.3.4 ([CGMS15, Theorem 2]).** For every \( \eta \in (0, 1/2) \), there exists \( \alpha > 0 \) and \( \beta > 0 \) such that for every \( k \) and \( \gamma \), it holds that every deterministic protocol \( \Pi \) that solves Agreement-Distillation\(^k_{\gamma, \eta}\) has communication complexity at least \( \alpha k - \beta \log(1/\gamma) \).

We note that while the agreement distillation problem is very similar to our (functional) agreement problem, there are some syntactic differences. We are considering pairs of functions with low communication complexity, whereas the agreement distillation problem considers arbitrary random sequences. Also, our output criterion is proximity to the input functions, whereas in the agreement distillation problem we need to produce high-entropy outputs. Finally, we want a lower bound for our agreement problem when Alice and Bob are allowed to share perfect randomness while the agreement distillation bound only holds for deterministic protocols. Nevertheless, we are able to reduce to the above setting of \cite{CGMS15} as we will see shortly.

Our proof of Theorem 5.3.2 uses the standard Chernoff-Hoeffding tail inequality for random variables that we include below.

**Proposition 5.3.5** (Chernoff bound; see e.g., \cite{MU05}). Let \( X = \sum_{i=1}^n X_i \) be a sum of independent identically distributed random variables \( X_1, \ldots, X_n \in \{0, 1\} \). Let \( \mu = \mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] \). It holds that for every \( \delta \in (0, 1) \),

\[
\Pr[X < (1 - \delta)\mu] \leq e^{-\delta^2 \mu/2}
\]
and

\[ \Pr[X > (1 + \delta)\mu] \leq e^{-\delta^2 \mu/3}, \]

and for \( a > 0 \),

\[ \Pr[X > \mu + a] \leq e^{-2a^2/n}. \]

**Proof of Theorem 5.3.2.** Let \( \mu \) be the uniform distribution on \( \mathcal{X} \times \mathcal{Y} \). We prove the theorem for \( \alpha/\beta < \delta/6 \), in which case we may assume that \( \gamma > \epsilon^{\delta|\mathcal{Y}|/6} \) since otherwise the right-hand side in the statement of Theorem 5.3.2 is non-positive.

Let \( \mathcal{F}_B \) denote the set of functions that depend only on Bob’s input, i.e., \( f \in \mathcal{F}_B \) if there exists \( f': \mathcal{Y} \to \{0, 1\} \) such that \( f(x, y) = f'(y) \) for all \( x, y \). Our family \( \mathcal{F} \) will be the subset of \( \mathcal{F}_B \times \mathcal{F}_B \) consisting of pairs of functions that are at most \( \delta \) apart (with respect to the uniform distribution on \( \mathcal{X} \times \mathcal{Y} \)), i.e.,

\[ \mathcal{F} \triangleq \{(f, g) \in \mathcal{F}_B \times \mathcal{F}_B \mid \delta_{\mu}(f, g) \leq \delta \}. \]

Note that the zero-error communication complexity of every function in the support of \( \mathcal{F} \) is zero since Bob can correctly compute its value without any information from Alice. Thus, \( \delta_{\mu}(\mathcal{F}) = \delta \) and \( CC_0(\mathcal{F}) = 0 \).\(^{14}\) So it remains to prove a lower bound on \( CC(\text{AGREE}_{\delta', \gamma}(\mathcal{F})) \).

We prove our lower bound by picking a distribution \( \mathcal{D}_\eta \) supported mostly on \( \mathcal{F} \) and giving a lower bound on \( CC(\text{AGREE}_{\delta', \gamma}(\mathcal{D}_\eta)) \). Let \( \eta = \delta/2 \). The distribution \( \mathcal{D}_\eta \) samples \((f, g)\) as follows. The function \( f \) is drawn uniformly at random from \( \mathcal{F}_B \). Since \( f \in \mathcal{F}_B \), there exists a function \( f': \mathcal{Y} \to \{0, 1\} \) such that \( f(x, y) = f'(y) \) for all \( x, y \). Then, \( g \) is chosen to be a “\( \eta \)-noisy copy” of \( f \). Namely, we define a function \( g': \mathcal{Y} \to \{0, 1\} \) such that for every \( y \in \mathcal{Y} \), \( g'(y) \) is chosen to be equal to \( f'(y) \) with probability \( 1 - \eta \) and equal to \( 1 - f'(y) \) with probability \( \eta \). Then, for every \( x \in \mathcal{X} \) and \( y \in \mathcal{Y} \), we set \( g(x, y) = g'(y) \).

By the Chernoff bound (Proposition 5.3.5), we have that

\[ \Pr_{(f, g) \sim \mathcal{D}_\eta} [\delta(f, g) > \delta] \leq e^{-\eta |\mathcal{Y}|/3} < \gamma. \]

So with probability at least \( 1 - \gamma \), the distribution \( \mathcal{D}_\eta \) draws elements from \( \mathcal{F} \). So, if a protocol solves \( \text{AGREE}_{\delta', \gamma}(\mathcal{F}) \), then if \( (f, g) \sim \mathcal{D}_\eta \), with probability at least \( (1 - \gamma) \cdot \gamma \), we would have that the function-pair \((f, g)\) belongs to \( \mathcal{F} \) and the protocol achieves agreement on a nearby function. We conclude that a protocol solving \( \text{AGREE}_{\delta', \gamma}(\mathcal{F}) \) is also a protocol solving \( \text{AGREE}_{\delta', \gamma - \gamma^2}(\mathcal{D}_\eta) \).

We thus need to show a lower bound on the communication complexity of \( \text{AGREE}_{\delta', \gamma - \gamma^2}(\mathcal{D}_\eta) \). We now note that since this is a distributional problem, by Yao’s min-max principle, if there is randomized protocol solving \( \text{AGREE}_{\delta', \gamma - \gamma^2}(\mathcal{D}_\eta) \), then there is also a deterministic protocol solving the same problem and with the same communication com-

\(^{14}\)Indeed, even the uncertain-communication complexity of \( \mathcal{F} \) is zero, further highlighting the lack of need of agreement to solve uncertain-communication problems.
plexity. Thus, it suffices to lower-bound the deterministic communication complexity of \( \text{AGREE}_{\delta', \gamma-\gamma^2}(D_\eta) \). Claim 5.3.6 below shows that any protocol solving this problem gives a deterministic protocol for AGREEMENT-DISTILLATION with \( k = \Omega_{\delta'}(|\mathcal{Y}|) \). Combining this with Lemma 5.3.4 gives us the desired lower bound on \( \text{CC}(\text{AGREE}_{\delta', \gamma-\gamma^2}(D_\eta)) \) and hence on \( \text{CC}(\text{AGREE}_{\delta', \gamma}(\mathcal{F})) \).

\[ \text{Claim 5.3.6.} \text{ Every protocol solving } \text{AGREE}_{\delta', \gamma}(D_\eta) \text{ is also a protocol for } \text{AGREEMENT-DISTILLATION}^{\text{sym}}_{\gamma, \eta} \text{ for } k = (1 - H_0(\delta')) \cdot |\mathcal{Y}|, \text{ where } \delta'' = \delta'(1 + o(1)) \text{ and } H_0(\cdot) \text{ is the binary entropy function given by } H_0(x) = -x \log x - (1 - x) \log(1 - x), \text{ where } o(1) \text{ denotes a function that goes to } 0 \text{ as } |\mathcal{Y}| \text{ grows.} \]

\[ \text{Proof.} \text{ Suppose Alice and Bob wish to solve } \text{AGREEMENT-DISTILLATION}^{\text{sym}}_{\gamma, \eta} \text{ with probability at least } \gamma. \text{ So it suffices to show that } H_\infty(h_A), H_\infty(h_B) \geq k. \]

\[ \text{The intuitive idea for establishing this is simple. In order to show that the min-entropy of } h_A \text{ (symmetrically, } h_B) \text{ is large, we need to argue that a given } h_A \text{ cannot be an output by a correct protocol for } \text{AGREE}_{\delta', \gamma}(D_\eta) \text{ with too high a probability. We expect this to be true because a given } h_A \text{ cannot be } \delta'-\text{close to too many input functions } f. \text{ In order to formally argue this, we define the real-valued function } h'_A : \mathcal{Y} \to [0, 1] \text{ as } h'_A(y) := E_{x \sim \mathcal{X}}[h_A(x, y)] \text{ for all } y \in \mathcal{Y}. \]

\[ \text{By the triangle inequality, we have that} \]

\[ \delta(h'_A, f') := E_{y \sim \mathcal{Y}}[|h'_A(y) - f'(y)|] \]

\[ = E_{y \sim \mathcal{Y}}[|E_{x \sim \mathcal{X}}[h_A(x, y) - f(x, y)]|] \]

\[ \leq E_{(x,y) \sim \mathcal{X} \times \mathcal{Y}}[|h_A(x, y) - f(x, y)|] \]

\[ = \delta(h_A, f) \]

\[ \leq \delta'. \]

\[ \text{We now define a “randomized rounding” of } h_A \text{ to be a random function } h' : \mathcal{Y} \to \{0, 1\} \text{ such that independently for each } y \in \mathcal{Y}, \text{ we have that } h'(y) = 1 \text{ with probability } h'_A(y), \text{ and } h'(y) = 0 \text{ with probability } 1 - h'_A(y). \]

\[ \text{Define } S \text{ to be the set of all Boolean-valued functions } \tilde{f} : \mathcal{Y} \to \{0, 1\} \text{ such that } \delta(h'_A, \tilde{f}') \leq \delta'. \text{ We now show that with probability } 1 - o(1) \text{ over the random choice of } h', \text{ at least a } 1 - o(1) \text{ fraction of the functions } \tilde{f}' \in S \text{ are such that } \delta(h', \tilde{f}') \leq \delta'(1 + o(1)). \]

\[ \text{To see this, note that for any fixed } \tilde{f}' \in S, \text{ we have that } E_{h}[\delta(h', \tilde{f}') = \delta(h'_A, \tilde{f}') \leq \delta', \text{ and hence by the Chernoff bound (Proposition 5.3.5)}, \]

\[ \Pr[\delta(h', \tilde{f}') > \delta'(1 + o(1))] \leq o(1). \]

\[ \text{This implies that} \]

\[ \mathbb{E}_{h'} \left[ \Pr_{\tilde{f}' \sim S} [\delta(h', \tilde{f}') > \delta'(1 + o(1))] \right] \leq o(1). \]

119
Thus, there exists a setting \( h' : \mathcal{Y} \to \{0, 1\} \) such that a \( 1 - o(1) \) fraction of the functions in \( S \) are within a distance of \( \delta'(1 + o(1)) \) from \( h' \). Thus,

\[
|S| \leq \left| \{ f' : \mathcal{Y} \to \{0, 1\} \mid \delta(h', f') \leq \delta'(1 + o(1)) \} \right| \cdot (1 + o(1)) \\
\leq 2^{H_b(\delta'(1 + o(1))|\mathcal{Y}|)} \cdot (1 + o(1)),
\]

where the last inequality follows from the fact that \( h' \) is a Boolean-valued function. Thus, we conclude that the right-hand side in Equation (5.3) is also an upper bound on the number of functions \( f : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \) such that \( f(x, y) = f'(y) \) for all \( x \in \mathcal{X}, y \in \mathcal{Y} \) for some function \( f' : \mathcal{Y} \to \{0, 1\} \) and that satisfy \( \delta(h_A, f) \leq \delta \). Since the probability of sampling any such \( f \) is equal to \( 2^{-|\mathcal{Y}|} \), we get that the probability of outputting any particular function \( h_A : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \) is at most \( 2^{-(1 - H_b(\delta'(1 + o(1)))|\mathcal{Y}|)} \). This means that \( H_{\infty}(h_A) \geq (1 - H_b(\delta'(1 + o(1))))|\mathcal{Y}| \). A similar lower bound applies to \( H_{\infty}(h_B) \). Thus, we have that the outputs of the protocol for \( \text{AGREE} \) solve \( \text{AGREEMENT-DISTILLATION}^k_{\gamma, \eta} \) with \( k = (1 - H_b(\delta'(1 + o(1))))|\mathcal{Y}| \).

\[\square\]

### 5.4 One-Way Uncertain Communication: Public-Coin Protocol

In this section, we prove Theorem 5.1.6 which we restate below (with a slight notational change compared to Section 5.1— we use \( \mathcal{X} \times \mathcal{Y} \) to denote the domain of the functions, as opposed to \( \{0, 1\}^n \times \{0, 1\}^n \)).

**Theorem 5.4.1** (Theorem 5.1.6 restated). There exists a positive constant \( c \) such that for all positive integers \( k \) and \( n \) and positive reals \( \epsilon, \delta \) and \( \theta \), for every distribution \( \mu \) over \( \{0, 1\}^n \times \{0, 1\}^n \), and every family \( \mathcal{F} \) of pairs of Boolean functions satisfying \( \delta_\mu(\mathcal{F}) \leq \delta \) and \( \text{owCC}_\epsilon^\mu(\mathcal{F}) \leq k \), it holds that

\[
\text{owPubCCU}^\mu_{\epsilon + 2\delta + \theta}(\mathcal{F}) \leq c \cdot \frac{k + \log \left( \frac{1}{\delta} \right)}{\theta^2} \cdot \left( 1 + \frac{I(X; Y)}{\theta^2} \right).
\]

### 5.4.1 Overview of Protocol

We start with a high-level description of the protocol. Let \( \mu \) be a distribution over an input space \( \mathcal{X} \times \mathcal{Y} \). For any function \( s : \mathcal{X} \times \mathcal{Y} \to \{0, 1\} \) and any \( x \in \mathcal{X} \), we define the restriction of \( s \) to \( x \) to be the function \( s_x : \mathcal{Y} \to \{0, 1\} \) given by \( s_x(y) = s(x, y) \) for any \( y \in \mathcal{Y} \). We will consider a pair \((X, Y)\) of random variables drawn from \( \mu \).

First, we consider the particular case of Theorem 5.1.6 where \( \mu \) is a product distribution, i.e., \( \mu = \mu_X \times \mu_Y \). Note that in this case, \( I(X; Y) = 0 \) in the right-hand side of Equation (5.1). We will handle the case of general (not necessarily product) distributions later on.

The general idea is that given inputs \((f, X)\), Alice can determine the restriction \( f_X \), and she will try to describe it to Bob. For most values \( x \in \mathcal{X} \), we have that \( f_x \) will be close (in \( \delta_{\mu_Y} \)-distance) to the function \( g_x \). Bob will try to use the (yet unspecified) description given
by Alice in order to determine some function $B$ that is close to $g_x$. If he succeeds in doing so, he can output $B(Y)$ which would equal $g_x(Y)$ with high probability over $Y$.

We next explain how Alice will describe $f_X$, and how Bob will determine some function $B$ that is close to $g_X$ based on Alice’s description. For the first part, we let Alice and Bob use shared randomness in order to sample $Y_1, \ldots, Y_m$, where the $Y_i$’s are drawn independently with $Y_i \sim \mu_Y$, and $m$ is a parameter to be chosen later. Alice’s description of $f_X$ will then be $(f_X(Y_1), \ldots, f_X(Y_m)) \in \{0, 1\}^m$. Thus, the length of the communication is $m$ bits and we need to show that setting $m$ to be roughly $O(k)$ suffices. Before we explain this, we first need to specify what Bob does with Alice’s message.

As a first cut, let us consider the following natural strategy: Bob picks an $\tilde{X} \in \mathcal{X}$ such that $g_{\tilde{X}}$ is close to $f_X$ on $Y_1, \ldots, Y_m$, and sets $B = g_{\tilde{X}}$. It is clear that if $\tilde{X} = X$, then $B = g_X = g_X$, and for every $y \in \mathcal{Y}$, we would have $B(y) = g_X(y)$. Moreover, if $\tilde{X}$ is such that $g_{\tilde{X}}$ is close to $g_X$ (which is itself close to $f_X$, for most values of $X$), then $B(Y)$ would now equal $g_X(Y)$ with high probability. It remains to deal with $\tilde{X}$ such that $g_{\tilde{X}}$ is far from $g_X$. Note that if we first fix any such $\tilde{X}$ and then sample $Y_1, \ldots, Y_m$, then with high probability, we would reveal that $g_{\tilde{X}}$ is far from $g_X$. This is because $g_X$ is close to $f_X$ (for most values of $X$), so $g_{\tilde{X}}$ should also be far from $f_X$. However, this idea alone cannot deal with all possible $\tilde{X}$ — using a naive union bound over all possible $\tilde{X} \in \mathcal{X}$ would require a failure probability of $1/|\mathcal{X}|$, which would itself require setting $m$ to be roughly $\log |\mathcal{X}|$. Indeed, smaller values of $m$ should not suffice since we have not yet used the fact that $\text{CC}^e(g) \leq k$ — but we do so next.

Suppose that $\Pi$ is a one-way protocol with $k$ bits of communication. Then, note that Alice’s message partitions $\mathcal{X}$ into $2^k$ sets, one corresponding to each message. Our modified strategy for Bob is to let him pick a representative from each set in this partition, and then set $B = g_{\tilde{X}}$ for an $\tilde{X}$ among the representatives for which $g_{\tilde{X}}$ and $f$ are the closest on the samples $Y_1, \ldots, Y_m$. A simple analysis shows that the $g_{\tilde{X}}$’s that lie inside the same set in this partition are close, and thus, if we pick $\tilde{X}$ to be the representative of the set containing $X$, then $g_{\tilde{X}}$ and $f_{\tilde{X}}$ will be close on the sampled points. For another representative, once again if $g_{\tilde{X}}$ is close to $g_X$, then $g_{\tilde{X}}(Y)$ will equal $g_X(Y)$ with high probability. For a representative $x'$ such that $g_{x'}$ is far from $g_X$ (which is itself close to $f_X$), we can proceed as in the previous paragraph, and now the union bound works out since the total number of representatives is only $2^k$.

We now turn to the case of general (not necessarily product) distributions. In this case, we would like to run the above protocol with $Y_1, Y_2, \ldots, Y_m$ sampled independently from $\mu_{Y|X}$ (instead of $\mu_Y$) where $x$ is the particular realization of Alice’s input. Note that Alice knows $x$ and hence knows the distribution $\mu_{Y|X}$. Unfortunately, Bob does not know $\mu_{Y|X}$; he only knows $\mu_Y$ as a “proxy” for $\mu_{Y|X}$. While Alice and Bob cannot jointly sample such $Y_i$’s without communicating (as in the product case), they can still run the rejection sampling protocol of [HJMR07] in order to agree on such samples while communicating at most $O(m \cdot I(X; Y))$

\footnote{We note that a similar idea was used in a somewhat different context by [BJKS02] (following on [KNR99]) in order to characterize one-way communication complexity of any function under product distributions in terms of its VC-dimension.}
bits (see Section 5.4.2 for more details).

The outline of the rest of this section is the following. In Section 5.4.2, we describe the properties of the correlated sampling procedure that we will use. In Section 5.4.3, we give the formal proof of Theorem 5.1.6.

5.4.2 Rejection Sampling

We start by recalling two standard notions from information theory. Given two distributions \( P \) and \( Q \), the KL divergence between \( P \) and \( Q \) is defined as

\[
D(P \parallel Q) \triangleq E_{u \sim P}[\log(P(u)/Q(u))].
\]

Given a joint distribution \( \mu \) of a pair \((X, Y)\) of random variables with \( \mu_X \) and \( \mu_Y \) being the marginals of \( \mu \) over \( X \) and \( Y \) respectively, the mutual information of \( X \) and \( Y \) is defined as

\[
I(X; Y) \triangleq D(\mu \parallel \mu_X \mu_Y).
\]

The following lemma summarizes the properties of the rejection sampling protocol of [HJMR07].

Lemma 5.4.2 (Rejection Sampling; [HJMR07]). Let \( P \) be a distribution known to Alice and \( Q \) be a distribution known to both Alice and Bob, with \( D(P \parallel Q) \) being finite. There exists a one-way public-coin protocol (with communication from Alice to Bob) such that at the end of the protocol, Alice and Bob output a sample from \( P \) such that the expected communication cost (over the public-randomness of the protocol) is at most \( D(P \parallel Q) + 2\log(D(P \parallel Q) + 1) + O(1) \) bits.

We will use the following corollary of Lemma 5.4.2.

Corollary 5.4.3. Let \( \mu \) be a distribution over \((X, Y)\) with marginal \( \mu_X \) over \( X \), and assume that \( \mu \) is known to both Alice and Bob. Fix \( \epsilon > 0 \) and let Alice be given a realization \( x \sim \mu_X \). There is a one-way public-coin protocol that uses at most

\[
O(m \cdot I(X; Y)/\epsilon) + O(1/\epsilon)
\]

bits of communication such that with probability at least \( 1 - \epsilon \) over the public coins of the protocol and the randomness of \( x \), Alice and Bob agree on \( m \) samples \( Y_1, Y_2, \ldots, Y_m \) i.i.d. \( \sim \mu_{Y|x} \) at the end of the protocol.

Proof. When \( x \) is Alice’s input, we can consider running the protocol in Lemma 5.4.2 on the distributions \( P \triangleq \prod_{i=1}^m \mu_{Y_i|x} \) and \( Q \triangleq \prod_{i=1}^m \mu_{Y_i} \). Note that each of \( P \) and \( Q \) is a distribution over tuples \((y_1, y_2, \ldots, y_m)\). Let \( \Pi \) be the resulting protocol transcript. The expected communication cost of \( \Pi \) is at most

\[
\mathbb{E}_{x \sim \mu_X} [O(D(P \parallel Q)) + O(1)] = O(\mathbb{E}_{x \sim \mu_X} [D(P \parallel Q)]) + O(1) = O(m \cdot I(X; Y)) + O(1),
\]

(5.4)
where the last equality follows from the fact that
\[
\mathbb{E}_{x \sim \mu_X}[D(P||Q)] = \mathbb{E}_{x \sim \mu_X} \left[ \mathbb{E}_{y_1, \ldots, y_m \mid x} \left[ \log \left( \frac{\prod_{i=1}^{m} \mu_{Y_i}(y_i)}{\prod_{i=1}^{m} \mu_{Y_i}(y_i)} \right) \right] \right]
\]
\[
= \sum_{i=1}^{m} \mathbb{E}_{x \sim \mu_X} \left[ \mathbb{E}_{y_1, \ldots, y_m \mid x} \left[ \log \left( \frac{\mu_{Y_i}(y_i)}{\mu_{Y_i}(y_i)} \right) \right] \right]
\]
\[
= \sum_{i=1}^{m} \mathbb{E}_{x \sim \mu_X} \left[ \log \left( \frac{\mu_{Y_i}(y_i)}{\mu_{Y_i}(y_i)} \right) \right]
\]
\[
= m \cdot I(X; Y).
\]

By Markov’s inequality applied to (Equation (5.4)), we get that with probability at least
\[1 - \epsilon,\]
the length of the transcript \(\Pi\) is at most
\[O(m \cdot I(X; Y)/\epsilon) + O(1/\epsilon) \text{ bits.}\]

The statement now follows. \(\square\)

### 5.4.3 Proof of Theorem 5.1.6

Recall that in the uncertain setting, Alice’s input is \((f, X)\) and Bob’s input is \((g, Y)\), where
\[(f, g) \in \mathcal{F}, (X, Y) \sim \mu\]
and \(\mathcal{F}\) is a family of pairs of Boolean functions satisfying \(\text{owCC}_\mu^\epsilon(\mathcal{F}) \leq k\) and \(\delta_\mu(\mathcal{F}) \leq \delta\). Let \(\Pi\) be the one-way protocol for \(g\) in the standard setting that shows that \(\text{owCC}_\mu^\epsilon(g) \leq k\). Note that \(\Pi\) can be described by an integer \(L \leq 2^k\) and functions \(\pi: \mathcal{X} \to [L]\) and \(\{B_i: \mathcal{Y} \to \{0, 1\}\}_{i \in [L]}\), such that Alice’s message on input \(X\) is \(\pi(X)\), and Bob’s output on message \(i\) from Alice and on input \(y\) is \(B_i(y)\). We use this notation below. We also set the parameter \(m = \Theta((k + \log(1/\theta))/\theta^2)\), which is chosen such that
\[2^k \cdot e^{-\theta^2 m/75} \leq 2\theta/5.\]

**Protocol.** The protocol \(\Pi'\) that we employ in the uncertain setting is described in Protocol 2. Roughly speaking, the protocol works as follows. First, Alice and Bob run the one-way rejection sampling procedure given by Corollary 5.4.3 in order to sample \(y_1, y_2, \ldots, y_m\) i.i.d. \(\sim \mu_{Y|x}\). Then, Alice sends the sequence \((f_x(y_1), \ldots, f_x(y_m))\) to Bob. Bob enumerates over \(i \in [L]\) and counts the fraction of \(z \in \{y_1, \ldots, y_m\}\) for which \(B_i(z) \neq f_x(z)\). For the index \(i\) which minimizes this fraction, Bob outputs \(B_i(y)\) and halts.
Protocol 2: The Uncertain-Communication Protocol II’

The Setting: Let \( \mu \) be a probability distribution over a message space \( \mathcal{X} \times \mathcal{Y} \). Alice and Bob are given functions \( f \) and \( g \), and inputs \( x \) and \( y \), respectively, where \((f, g) \in \mathcal{F} \) and \((x, y) \) are realizations of the random pair \( (X, Y) \sim \mu \).

The Protocol:

1. Alice and Bob run one-way rejection sampling with error parameter set to \( (\theta/10)^2 \) in order to sample \( m \) values \( Z = \{y_1, y_2, \ldots, y_m\} \subseteq \mathcal{Y} \) each sampled independently according to \( \mu_{Y|x} \).
2. Alice sends \( \{f_x(y_i)\}_{i \in [m]} \) to Bob.
3. For every \( i \in [L] \), Bob computes \( \text{err}_i \triangleq \frac{1}{m} \sum_{j=1}^{m} 1(B_i(y_j) \neq f_x(y_j)) \). Let \( i_{\text{min}} \triangleq \arg\min_{i \in [L]} \{\text{err}_i\} \). Bob outputs \( B_{i_{\text{min}}}(y) \) and halts.

Analysis. Observe that by Corollary 5.4.3, the rejection sampling procedure requires \( O(m \cdot I(X; Y)/\theta^2 + 1/\theta^2) \) bits of communication. Thus, the total communication of our protocol is at most

\[
O(m \cdot I(X; Y)/\theta^2 + 1/\theta^2) + m \leq \frac{c(k + \log \left(\frac{1}{\theta}\right))}{\theta^2} \cdot \left(1 + \frac{I(X; Y)}{\theta^2}\right)
\]

bits for some absolute constant \( c \), as claimed. The next lemma establishes the correctness of the protocol.

Lemma 5.4.4. \( \Pr_{\text{II’}(x, y) \sim \mu} [B_{i_{\text{min}}}(y) \neq g(x, y)] \leq \epsilon + 2\delta + \theta \), where the probability is over both the internal randomness of the protocol II’ and over the randomness of the input-pair \((x, y)\).

Proof. We start with some notation. For \( x \in \mathcal{X} \), let \( \delta_x \triangleq \delta_{\mu_{Y|x}} (f_x, g_x) \) and let \( \epsilon_x \triangleq \delta_{\mu_{Y|x}} (g_x, B_{\pi(x)}) \). Note that by definition, \( \delta = \mathbb{E}_{x \sim \mu_X}[\delta_x] \) and \( \epsilon = \mathbb{E}_{x \sim \mu_X}[\epsilon_x] \). For \( i \in [L] \), let \( \gamma_{i,x} \triangleq \delta_{\mu_{Y|x}} (f_x, B_i) \). Recall the description of the (given) deterministic protocol II by the positive integer \( L \leq 2^k \) and functions \( \pi: \mathcal{X} \rightarrow [L] \) and \( \{B_i: \mathcal{Y} \rightarrow \{0, 1\}\}_{i \in [L]} \), such that Alice’s message on input \( x \) is \( \pi(x) \), and Bob’s output on message \( i \) from Alice and on input \( y \) is \( B_i(y) \). Note that by the triangle inequality,

\[
\gamma_{\pi(x), x} = \delta_{\mu_{Y|x}} (f_x, B_{\pi(x)}) \leq \delta_x + \epsilon_x. \quad (5.5)
\]

In what follows, we will analyze the probability that \( B_{i_{\text{min}}}(y) \neq g(x, y) \) by analyzing the estimate \( \text{err}_i \) and the index \( i_{\text{min}} \) computed in Protocol 2. Note that in this protocol, both \( \text{err}_i \) and \( i_{\text{min}} \) are functions of \( x \) and the computed \( \text{err}_i = \text{err}_i(x) \) attempts to estimate \( \gamma_{i,x} \).

Note that Corollary 5.4.3 guarantees that rejection sampling succeeds with probability at least \( 1 - \theta^2/100 \). Henceforth, we condition on the event that rejection sampling succeeds.
(we will account for the event where this does not happen at the end). By the Chernoff bound (Proposition 5.3.5), and using the definition of $err_i$ in Protocol 2 and the fact that

$$\mathbb{E}[1(B_i(y_j) \neq f_x(y_j))] = \gamma_{i,x},$$

we have for every $x$ and $i \in [L]$

$$\Pr_{y_1, \ldots, y_m \sim \mu_Y | x} \left[ |\gamma_{i,x} - err_i| > \frac{\theta}{5} \right] \leq e^{-\frac{\theta^2 m}{10}}.$$

By a union bound, we have for every $x \in X$,

$$\Pr_{y_1, \ldots, y_m \sim \mu_Y | x} \left[ \exists i \in [L] \text{ s.t. } |\gamma_{i,x} - err_i| > \frac{\theta}{5} \right] \leq L \cdot e^{-\frac{\theta^2 m}{10}} \leq \frac{2\theta}{5},$$

where the last inequality follows from our choice of $m = \Theta((k + \log(1/\theta))/\theta^2)$.

Now assume that for all $i \in [L]$, we have that $|\gamma_{i,x} - err_i| \leq \theta/5$, which we refer to below as the “Good Event.” Then, for $i = i_{\min}$, we have that

$$\gamma_{i_{\min}, x} \leq err_{i_{\min}} + \theta/5 \quad \text{(since we assumed the Good Event)}$$

$$\leq err_{\pi(x)} + \theta/5 \quad \text{(by definition of } i_{\min})$$

$$\leq \gamma_{\pi(x), x} + 2\theta/5 \quad \text{(since we assumed the Good Event)}$$

$$\leq \delta_x + \epsilon_x + 2\theta/5. \quad \text{(by Equation (5.5))}$$

Let $W \subseteq \{0,1\}^n$ be the set of all $x$ for which rejection sampling succeeds with probability at least $1 - \theta/10$ (over the internal randomness of the protocol). By Corollary 5.4.3 and an averaging argument, $\Pr_{x \sim \mu_X} [x \notin W] \leq \theta/10$. Denoting by $\mu_X | x \in W$ the conditional probability distribution of $x \sim \mu_X$ conditioned on the event that $x \in W$, we thus get that

$$\Pr_{x,y \sim \mu} [B_{i_{\min}}(y) \neq f(x, y)]$$

$$\leq \Pr_{x \sim \mu_X} [x \in W] \cdot \mathbb{E}_{x \sim \mu_X | x \in W} \left[ \Pr_{y \sim \mu_Y | x} [B_{i_{\min}}(y) \neq f(x, y)] \right] + \frac{\theta}{10}$$

$$\leq \Pr_{x \sim \mu_X} [x \in W] \cdot \mathbb{E}_{x \sim \mu_X | x \in W} \left[ \Pr_{y_1, \ldots, y_m, y \sim \mu_Y | x} [B_{i_{\min}}(y) \neq f(x, y)] \right] + \theta/5$$

$$= \Pr_{x \sim \mu_X} [x \in W] \cdot \mathbb{E}_{x \sim \mu_X | x \in W} [\gamma_{i_{\min}, x}] + \theta/5$$

$$\leq \Pr_{x \sim \mu_X} [x \in W] \cdot \mathbb{E}_{x \sim \mu_X | x \in W} [\delta_x + \epsilon_x + 2\theta/5] + 3\theta/5$$

$$\leq \mathbb{E}_{x \sim \mu_X} [\delta_x + \epsilon_x] + \theta$$

$$= \delta + \epsilon + \theta,$$

where the third inequality follows from the fact that the Good Event occurs with probability
at least $1 - 2\theta/5$, and from the corresponding upper bound on $\gamma_{\text{min}}$. The other inequalities above follow from the definition of the set $W$ and the fact that $\Pr_{x \sim \mathcal{X}}[x \notin W] \leq \theta/10$. Finally, since $\delta(f, g) \leq \delta$, we have that Bob’s output does not equal $g(x, y)$ (which is the desired output) with probability at most $\epsilon + 2\delta + \theta$.  

\section{One-Way Uncertain Communication: Imperfectly Shared Randomness Protocol}

In this section, we prove Theorem 5.1.9 which we start by recalling.

\textbf{Theorem 5.5.1} (Theorem 5.1.9 restated). Let $\rho \in (0, 1]$ and $\mu$ be a product distribution. Let $\mathcal{F}$ consist of pairs $(f, g)$ of functions with $\Delta_{\mu}(f, g) \leq \delta$, and $\text{owCC}_\mu(f), \text{owCC}_\mu(g) \leq k$. Then, for every positive $\theta$, $\text{owIsrCC}_{\epsilon + 2\delta + \theta}(\mathcal{F}) \leq O(\theta k/\rho^2)$.

In order to prove Theorem 5.1.9, we start by defining a communication problem that will be useful to us (a similar definition is used in [CGMS15]).

\textbf{Definition 5.5.2} (Gap Inner Product; $\text{GIP}_{c,s}$). Let $-1 \leq s < c \leq 1$. Let Alice be given a vector $u \in \{\pm 1\}^d$ and Bob be given a vector $v \in \{\pm 1\}^d$. The goal is for Alice and Bob to distinguish the case where $E_{i \in [d]}[u_i v_i] \geq c$ from the case where $E_{i \in [d]}[u_i v_i] \leq s$.

We let ISR$_\rho$ denote the source of imperfect shared randomness where Alice gets a string $r$ of independent uniform-random bits and Bob gets a string $r'$ of bits obtained by independently flipping each coordinate of $r$ with probability $(1 - \rho)/2$. For a function $f$, we denote by $\text{owIsrCC}_{\epsilon,\rho}(f)$ the minimum cost of a one-way protocol that has access to ISR$_\rho$ and that on input $(x, y)$ in the domain of $f$, outputs $f(x, y)$ with probability at least $1 - \epsilon$, where the probability is over the randomness of ISR$_\rho$.

The following theorem –which upper bounds the communication complexity with ISR$_\rho$ of $\text{GIP}_{c,s}$— was proved by [CGMS15] using a locality-sensitive-hashing based protocol.

\textbf{Theorem 5.5.3} ([CGMS15]). Let $\rho \in (0, 1]$. Then, $\text{owIsrCC}_{\epsilon,\rho}(\text{GIP}_{c,s}) = O((c-s)^{-2}\rho^{-2}\log(1/\epsilon))$ via a protocol where Alice’s message depends only on her input, her part of the randomness, the values of $\rho$ and $\epsilon$ and the difference $c - s$.

Theorem 5.5.3 can be used in order to estimate the weighted inner product of two vectors up to an arbitrary additive accuracy, and when the weighting is done according to an arbitrary distribution on coordinates that is known to both Alice and Bob.

\textbf{Lemma 5.5.4}. Let $t, d \in \mathbb{N}$ and $P$ be a distribution over $[d]$ that is known to both Alice and Bob. Let Alice be given a vector $u \in \{\pm 1\}^d$ and Bob be given $t$ vectors $v^{(1)}, v^{(2)}, \ldots, v^{(t)} \in \{\pm 1\}^d$. Let $\theta > 0$ and $\rho \in (0, 1]$ be given. Then, there exists a one-way protocol with communication cost $O(\theta^{-2}\rho^{-2}\log(t/\theta))$ bits such that, with probability $1 - \theta$, for every $j \in [t]$, Bob computes $E_{i \sim P}[u_i v^{(j)}_i]$ up to an additive accuracy of $\theta$.  

126
Proof. First, note that we can reduce the case of general distributions \( P \) on \([d]\) to the case of the uniform distribution on \([d']\) for some integer \( d' \in \mathbb{N} \), by having Alice and Bob repeat coordinate \( i \in [d] \) of their vectors a number of times proportional to \( P(i) \). More precisely, we can assume without loss of generality that \( P(i) \) is a rational number for each \( i \in [d] \) (because of the density of the rationals in the reals), and then have Alice and Bob repeat coordinate \( i \) a number of times equal to \( \ell \cdot P(i) \) where \( \ell \) is the least-common multiple of the denominators in \( \{P(i): i \in [d]\} \).

Henceforth, we assume that \( P \) is the uniform distribution on \([d']\). Alice sends a message of the protocol for GIP in Theorem 5.5.3 with parameters \( c - s = \theta \) and \( \epsilon = \theta^2/t \). Bob then divides the interval \([-1, +1]\) into \( 2/\theta \) sub-intervals, each of length \( \theta \). Then, he completes the protocol for GIP in Theorem 5.5.3 on Alice’s message, for each subinterval and for each of his vectors \( v(1), v(2), \ldots, v(t) \). For each fixed \( j \in [t] \), by a union bound over the \( 2/\theta \) sub-intervals, we get that with probability \( 1 - \theta/t \), Bob can deduce the value of \( \mathbb{E}_{i \sim P}[u_i v_i(j)] \) up to additive accuracy \( \theta \). Another union bound over all \( t \) vectors of Bob implies that with probability \( 1 - \theta \), for each \( j \in [t] \), he computes the value of \( \mathbb{E}_{i \sim P}[u_i v_i(j)] \) up to additive accuracy \( \theta \).

Moreover, by our setting of \( c - s = \theta \) and \( \epsilon = \theta^2/t \), we get that the communication cost of the protocol is \( O(\theta^{-2}\rho^{-2}\log(t/\theta)) \) bits.

We are now ready to prove Theorem 5.1.9.

Proof of Theorem 5.1.9. Let \( \rho \in [0, 1] \) and \( \mu \) be a product distribution. Consider a pair \((f, g) \in \mathcal{F}\). Since \( \text{owCC}_\mu^\rho(g) \leq k \), there exist an integer \( L \leq 2^k \) and (deterministic) functions \( \pi: X \to [L] \) and \( \{B_i: Y \to \{0, 1\}\}_{i \in [L]} \), such that Alice’s message on input \( x \) is \( \pi(x) \), and Bob’s output on message \( i \) from Alice and on input \( y \) is \( B_i(y) \). For every \( x \in X \), we define the function \( f_x: Y \to \{0, 1\} \) as \( f_x(y) \triangleq f(x, y) \) for all \( y \in Y \). We also denote \( \tilde{f}_x(y) \triangleq (-1)^{f_x(y)} \), and similarly \( \tilde{B}_i(y) \triangleq (-1)^{B_i(y)} \). The operation of the protocol is given in Protocol 3.

**Protocol 3** The Uncertain-Communication Protocol with ISR\(\rho\)

**Setting:** Let \( \mu \) be a product distribution over a message space \( X \times Y \). Alice and Bob are given functions \( f \) and \( g \), and inputs \( x \) and \( y \), respectively, where \( \Delta_\mu(f, g) \leq \delta \), \( \text{owCC}_\mu^\rho(f) \), \( \text{owCC}_\mu^\rho(g) \leq k \) and \((x, y) \sim \mu\).

**Protocol:**

1. Alice and Bob run the protocol in Lemma 5.5.4 with \( t = L, d = |Y|, \) \( P = \mu_Y \), accuracy \( \theta/3, u = f_x, v(j) = \tilde{B}_j \) for every \( j \in [L] \).

2. For every \( j \in [L] \), let \( \text{agr}_j \) denote Bob’s estimate of \( \mathbb{E}_{y \sim \mu_Y} [\tilde{f}_x(y) \tilde{B}_j(y)] \).

3. Bob determines \( j_{\text{max}} \triangleq \text{argmax}_{j \in [L]} \{\text{agr}_j\} \) and outputs \( B_{j_{\text{max}}}(y) \) and halts.

The same argument as in Lemma 5.4.4 (specialized to product distributions) then implies
that the probability that \(B_{i_{\max}}(y)\) is not equal to \(g(x, y)\) is at most \(\epsilon + 2\delta + \theta\). By Lemma 5.5.4 and our setting of \(t = L \leq 2^k\), we get that the communication cost of Protocol 3 is \(O_\theta(k/\rho^2)\) bits, as desired. □

### 5.6 Lower Bound on Public-Coin Protocols

In this section, we prove Theorem 5.1.3, or rather a slight strengthening of this theorem as stated below.

**Theorem 5.6.1.** There exist absolute constants \(\alpha > 0\) and \(\beta < \infty\) such that for every positive integer \(n\), every \(\delta \in (0, 1)\), and every \(\epsilon < 1/2 - 2^{-\beta \sqrt{n}}\) the following holds: There exists a distribution \(\mu\) supported on \(\{0, 1\}^n\) and a function class \(\mathcal{F}\) satisfying \(\delta_\mu(\mathcal{F}) \leq \delta\) and \(\text{owCC}_0^\epsilon(\mathcal{F}) \leq 1\) such that

\[
\text{PubCCU}_\epsilon(\mathcal{F}) \geq \alpha \sqrt{\delta n} - \log \left(\frac{2}{1/2 - \epsilon}\right).
\]

Note that Theorem 5.1.3 is the special case of Theorem 5.6.1 where \(\delta\) and \(\epsilon\) are absolute constants.

To prove Theorem 5.6.1, we start by defining the class of function pairs and distributions that will be used. Consider the parity functions on subsets of bits of the string \(x \oplus y \in \{0, 1\}^n\) (which is the coordinate-wise XOR of the strings \(x, y \in \{0, 1\}^n\)). Specifically, for every \(S \subseteq [n]\), let \(\chi_S : \{0, 1\}^n \to \{0, 1\}\) be defined by \(\chi_S(x) = \bigoplus_{i \in S} x_i\) for all \(x \in \{0, 1\}^n\), and let \(f_S : \{0, 1\}^n \times \{0, 1\}^n \to \{0, 1\}\) be defined as

\[
f_S(x, y) \triangleq \chi_S(x \oplus y) = \bigoplus_{i \in S} (x_i \oplus y_i).
\]

Let \(q = q(n) > 0\) and define the class of pairs of Boolean functions

\[
\mathcal{F}_q \triangleq \{(f_S, f_T) : |S \triangle T| \leq q \cdot n\}. \quad (5.6)
\]

Next, we define a probability distribution \(\mu_p\) on \(\{0, 1\}^n \times \{0, 1\}^n\) where \(p = p(n)\). We do so by giving a procedure to sample according to \(\mu_p\). To sample a pair \((X, Y) \sim \mu_p\), we draw \(X \sim \{0, 1\}^n\) (i.e., we draw \(X\) uniformly at random from \(\{0, 1\}^n\)) and let \(Y\) be a \(p\)-noisy copy of \(X\), i.e., \(Y \sim N_p(X)\). Here, for any \(x \in \{0, 1\}^n\), \(N_p(x)\) is the distribution on \(\{0, 1\}^n\) that outputs \(Y \in \{0, 1\}^n\) such that, independently, for each \(i \in [n]\), \(Y_i = 1 - x_i\) with probability \(p\), and \(Y_i = x_i\) with probability \(1 - p\). In other words, \(\mu_p(x, y) = 2^{-n} \cdot p^{|x \oplus y|} \cdot (1 - p)^{n - |x \oplus y|}\) for every \((x, y) \in \{0, 1\}^n \times \{0, 1\}^n\) where the notation \(|z|\) stands for the Hamming weight of \(z\), for \(z \in \{0, 1\}^n\).

We will prove Lemmas 5.6.2 and 5.6.3 below about the function class \(\mathcal{F}_q\) and the distribution \(\mu_p\). In words, Lemma 5.6.2 says that every pair of functions in \(\mathcal{F}_q\) are \((pqn)\)-close in \(\delta_{\mu_p}\)-distance, and every function in \(\mathcal{F}_q\) has a one-way zero-error certain-communication protocol with a single bit of communication. Lemma 5.6.3 lower-bounds the uncertain-communication complexity of \(\mathcal{F}_q\) under distribution \(\mu_p\).
Lemma 5.6.2. For every positive integer \( n \) and every \( p, q \in [0, 1] \), we have that \( \text{owCC}^μ_0(F_q) \leq 1 \) and \( \delta_{μ_p}(F_q) \leq pqn \).

Lemma 5.6.3. There exist constants \( γ, τ > 0 \) such that for every positive integer \( n \), every \( p \in (0, 1/2) \), \( q \in (0, 1) \) and \( ε < 1/2 \), it holds that:

\[
\text{PubCC}^μ_p(F_q) ≥ γ \cdot \min\{p \cdot n, (q/2) \cdot n\} - \log \left( \frac{1}{1/2 - (ε + η)} \right),
\]

where \( η = 2^{-τ \cdot q \cdot n} \).

Note that applying Lemmas 5.6.2 and 5.6.3 with \( F = F_q \), \( μ = μ_p \) and \( p = q = \sqrt{δ/n} \) (where \( δ > 0 \)) implies Theorem 5.6.1.

In Section 5.6.1 below, we prove Lemma 5.6.2 which follows from two simple propositions. The main part of the rest of this section is dedicated to the proof of Lemma 5.6.3. The idea behind the proof of Lemma 5.6.3 is to reduce the problem of computing \( F_q \) under \( μ_p \) with uncertainty, from the problem of computing a related function in the standard distributional communication complexity model (i.e., without uncertainty) under a related distribution. We then use the discrepancy method to prove a lower bound on the communication complexity of the new problem. This task itself reduces to upper-bounding the spectral norm of a specific communication matrix. The choice of our underlying distribution then implies a tensor structure for this matrix, which reduces the spectral norm computation to bounding the largest singular value of an explicit family of \( 4 \times 4 \) matrices.

We point out that our lower bound in Lemma 5.6.3 is essentially tight up to a logarithmic factor. Namely, one can show using a simple one-way hashing protocol that for any constant \( ε > 0 \), \( \text{owPubCC}^μ_r(F_q) \leq O(r \cdot \log r) \) with \( r \triangleq \min\{p \cdot n, (q/2) \cdot n\} \). More precisely, let us first assume that \( p \leq (q/2) \), in which case \( r = p \cdot n \). Then, with very high probability, \( x \) and \( y \) are within a Hamming distance of \( 2 \cdot r \). Thus, Bob can learn \( x \oplus y \) (and thus deduce \( x \)) if Alice sends him a (one-way) message of \( O(r \cdot \log r) \) bits. Specifically, when \( r = \Theta(\sqrt{n}) \), it can be seen that the one-way protocol for the “\( (2 \cdot r) \)-Hamming distance problem” (see e.g., [HSZZ06] and [BBG14]) reveals to Bob the coordinate-wise XOR of \( x \) and \( y \). Hence, Bob can deduce \( x \) and output \( χ_T(x \oplus y) \) in order to solve the uncertain problem in Lemma 5.6.3.

The case where \( p > (q/2) \) is similar, except that an “\( r \)-Hamming distance protocol” is now applied to the pair \( (S, T) \) (instead of the pair \( (x, y) \)); this would allow Bob to deduce \( S \) and, upon receiving the bit \( χ_S(y) \) from Alice, he can output \( χ_S(x \oplus y) = χ_S(x) \oplus χ_S(y) \).

### 5.6.1 Proof of Lemma 5.6.2

Lemma 5.6.2 follows from Proposition 5.6.4 and Proposition 5.6.5 below. We first show that every two functions in \( F_q \) are close under the distribution \( μ_p \).

**Proposition 5.6.4.** For every \( (f, g) \in F_q \), it holds that \( \delta_{μ_p}(f, g) \leq pqn \).
Proof. Any pair of functions \((f, g) \in \mathcal{F}_q\) is of the form \(f = f_S\) and \(g = f_T\) with \(|S \triangle T| \leq q \cdot n\). Hence,

\[
\Pr_{(x, y) \sim \mu} [f(x, y) \neq g(x, y)] = \Pr_{(x, y) \sim \mu} [\chi_{S \triangle T}(x \oplus y) = 1] \\
\leq 1 - (1 - p)^{|S \triangle T|} \\
\leq 1 - (1 - p)^{qn} \\
\leq pqn.
\]

Next, we show that there is a simple one-way communication protocol that allows Alice and Bob to compute \(f_S\) (for any \(S \subseteq [n]\)) with just a single bit of communication.

**Proposition 5.6.5.** \(\text{owCC}(f_S) = 1\).

**Proof.** Recall that \(f_S(x, y) = \oplus_{i \in S} (x_i \oplus y_i)\). We write this as \(f_S(x, y) = (\oplus_{i \in S} (x_i)) \oplus (\oplus_{i \in S} (y_i))\). This leads to the simple one-way protocol where Alice computes \(b = \oplus_{i \in S} (x_i)\) and sends the single bit result of the computation to Bob. Bob can now compute \(b \oplus (\oplus_{i \in S} (y_i)) = f_S(x, y)\) to obtain the value of \(f_S\) (with zero error).

5.6.2 Proof of Lemma 5.6.3

In order to lower-bound \(\text{PubCCU}_{\epsilon}^{\mu_p}(\mathcal{F}_q)\), we define a certain-communication problem in the distributional setting that can be reduced to the problem of computing \(\mathcal{F}_q\) in the uncertain setting. The lower bound in Lemma 5.6.3 is then obtained by proving a lower bound on the communication complexity of the new problem which is defined as follows:

- **Inputs:** Alice’s input is a pair \((S, x)\) where \(S \subseteq [n]\) and \(x \in \{0, 1\}^n\). Bob’s input is a pair \((T, y)\) such that \(T \subseteq [n]\) and \(y \in \{0, 1\}^n\).

- **Function:** The goal is to compute the function \(F\) given by

\[
F((S, x), (T, y)) \triangleq f_T(x, y) = \chi_T(x \oplus y).
\]

- **Distribution:** Let \(\mathcal{D}_q\) be a distribution on pairs of subsets \((S, T)\) of \([n]\) defined by the following sampling procedure. To sample \((S, T) \sim \mathcal{D}_q\), we pick a subset \(S \subseteq [n]\) uniformly at random, and we then sample \(T\) by letting its 0/1 indicator vector be a \((q/2)\)-noisy copy of the 0/1 indicator vector of \(S\). The joint distribution on the inputs of Alice and Bob is then described by \(\nu_{p, q} = \mathcal{D}_q \otimes \mu_p\); we sample \((x, y) \sim \mu_p\) and independently sample \((S, T) \sim \mathcal{D}_q\).

Proposition 5.6.6 below — which follows from a simple Chernoff bound — shows that a protocol computing \(\mathcal{F}_q\) under \(\mu_p\) can also be used to compute the function \(F\) in the standard distributional model with \(((S, x), (T, y)) \sim \nu_{p, q}\), and with the same amount of communication.
Proposition 5.6.6. There exists \( \tau > 0 \) such that for every \( \epsilon < 1/2 \), it holds that \( \text{PubCCU}^{\mu_p}_\epsilon(F_q) \geq \text{CC}^{\nu_{p,q}}_{\epsilon + \eta}(F) \) with \( \eta = 2^{-\tau \cdot q \cdot n} \).

Proof. Since \( ((S, x), (T, y)) \sim \nu_{p,q} \), we have that \( (x, y) \sim \mu_p \) and \( (S, T) \sim \mathcal{D}_q \). Thus, it suffices to show that for \( (S, T) \sim \mathcal{D}_q \), it holds that \( |S \Delta T| \leq q \cdot n \) with probability at least \( 1 - \eta \), where \( \eta = 2^{-\tau \cdot q \cdot n} \) for some universal constant \( \tau > 0 \). This follows from the definition of \( \mathcal{D}_q \), the Chernoff bound (Proposition 5.3.5) and the fact that \( E_{(S, T) \sim \mathcal{D}_q}(|S \Delta T|) = (q \cdot n)/2 \). \( \square \)

In the rest of this section, we will prove the following lower bound on \( \text{CC}^{\nu_{p,q}}_{\epsilon}(F) \), which along with Proposition 5.6.6, implies Lemma 5.6.3:

Lemma 5.6.7. There exists \( \gamma > 0 \) such that for every positive integer \( n \), every \( p \in (0, 1/2) \), \( q \in (0, 1) \) and \( \epsilon < 1/2 \), we have that

\[
\text{CC}^{\nu_{p,q}}_{\epsilon}(F) \geq \gamma \cdot \min\{p \cdot n, (q/2) \cdot n\} - \log \left( \frac{1}{1/2 - \epsilon} \right).
\]

We first state and prove a proposition that allows us to eliminate one of the two parameters \( p \) and \( q \).

Proposition 5.6.8. For every positive integer \( n \), every \( p \in (0, 1/2) \), \( q \in (0, 1) \) and \( \epsilon < 1/2 \), we have that \( \text{CC}^{\nu_{p,q}}_{\epsilon}(F) \geq \text{CC}^{\nu_{r,2r}}_{\epsilon}(F) \), where \( r \triangleq \min(p, q/2) \).

Proof. We use the fact that Alice and Bob can perturb their inputs (using private randomness) to reduce the correlations among them. Specifically, we use the fact that if \( y \) is a \( p \)-noisy copy of \( x \) and \( z \) is a \( \eta \)-noisy copy of \( y \), then \( z \) is an \( (p(1 - \eta) + \eta(1 - p)) \)-noisy copy of \( x \). Below, we show how to use this formally in a reduction.

Suppose \( ((S, x), (T, y)) \sim \nu_{r,2r} \) and Alice has \( (S, x) \) and Bob has \( (T, y) \) and the goal is to compute \( F((S, x), (T, y)) = \chi_T(x \oplus y) \). Suppose \( \Pi \) is a protocol with communication complexity \( k \) and that \( \epsilon \)-computes \( F \) on \( \nu_{p,q} \).

If \( q/2 > p = r \), then Alice samples a subset \( S' \eta \)-noisily from the set \( S \) for \( \eta = (q/2 - r)/(1 - 2r) \), so that \( (S', T) \sim \mathcal{D}_q \). Alice and Bob can now compute \( \Pi((S', x), (T, y)) \) using \( k \) bits of communication. By the correctness of \( \Pi \), we have that their output disagrees with \( F((S', x), (T, Y)) \) with probability at most \( \epsilon \). But then we have \( F((S, x), (T, y)) = F((S', x), (T, y)) \) since \( F \) does not depend on \( S \), and so Bob can simply output the output of \( \Pi \) to get a protocol that \( \epsilon \)-computes \( F \) on \( \nu_{r,2r} \).

Now we turn to the case that \( p \geq q/2 = r \). In this case, Bob samples \( y' \eta \)-noisily from \( y \), for \( \eta = (p - r)/(1 - 2r) \), to get \( y' \) which is an \( r \)-noisy copy of \( x \). By simulating \( \Pi((S, x), (T, y')) \), Bob can \( \epsilon \)-compute \( \chi_T(x \oplus y') \). Now using the fact that \( \chi_T(x \oplus y) = \chi_T(x \oplus y') \oplus \chi_T(y' \oplus y) \), we have that if Bob outputs \( \Pi((S, x), (T, y')) \oplus \chi_T(y' \oplus y) \), then he gets a protocol that is correct with probability at least \( 1 - \epsilon \).

Thus, in either case, \( \Pi \) can be converted to a protocol with the same communication and that \( \epsilon \)-computes \( F \) on \( \nu_{r,2r} \). \( \square \)

So in order to prove Lemma 5.6.7, we will set \( q = 2p \) and prove a lower bound of \( \gamma \cdot p \cdot n - \log(1/(1/2 - \epsilon)) \) on the communication complexity. So henceforth, we denote
The proof will use the \textit{discrepancy bound} which is a well-known method for proving lower bounds on distributional communication complexity in the standard model.

**Definition 5.6.9** (Discrepancy \cite[Definition 3.27]{KN97}). Let $F$ and $\nu_p$ be as above and let $R$ be any rectangle (i.e., any set of the form $R = C \times D$ where $C, D \subseteq \{0,1\}^n$). Denote

\[
\text{Disc}_{\nu_p}(R, F) \triangleq \left| \Pr_{\nu_p} \left[ \left( F((S, x), (T, y)) = 0 \text{ and } ((S, x), (T, y)) \in R \right) \right] - \Pr_{\nu_p} \left[ \left( F((S, x), (T, y)) = 1 \text{ and } ((S, x), (T, y)) \in R \right) \right] \right|.
\]

The discrepancy of $F$ according to $\nu_p$ is

\[
\text{Disc}_{\nu_p}(F) \triangleq \max_R \text{Disc}_{\nu_p}(R, F),
\]

where the maximum is over all rectangles $R$.

The next known proposition relates distributional communication complexity to discrepancy.

**Proposition 5.6.10** \cite[Proposition 3.28]{KN97}. For every $\epsilon < 1/2$, it holds that $\text{CC}_{\nu_p}^\epsilon(F) \geq \log((1 - 2\epsilon)/\text{Disc}_{\nu_p}(F))$.

We will prove the following lemma.

**Lemma 5.6.11.** $\text{Disc}_{\nu_p}(F) \leq 2^{-\gamma p n}$ for some absolute constant $\gamma > 0$.

Note that Lemma 5.6.11, Proposition 5.6.10 and Proposition 5.6.8 put together immediately imply Lemma 5.6.7. The proof of Lemma 5.6.11 uses some standard facts about the spectral properties of matrices and their tensor powers that we next recall. Let $A \in \mathbb{R}^{d \times d}$ be a real square matrix. Then, $v \in \mathbb{R}^d$ is said to be an eigenvector of $A$ with eigenvalue $\lambda \in \mathbb{R}$ if $Av = \lambda v$. If $A$ is furthermore (symmetric) positive semi-definite, then all its eigenvalues are real and non-negative. We can now define the spectral norm of a (not necessarily symmetric) matrix.

**Definition 5.6.12.** The spectral norm of a matrix $A \in \mathbb{R}^{d \times d}$ is given by $\|A\| \triangleq \sqrt{\lambda_{\text{max}}(A^T A)}$, where $\lambda_{\text{max}}(A^T A)$ is the largest eigenvalue of $A^T A$.

Also, recall that given a matrix $A \in \mathbb{R}^{d \times d}$ and a positive integer $t$, the tensor power matrix $A^{\otimes t} \in \mathbb{R}^{d^t \times d^t}$ is defined by $(A^{\otimes t})_{(i_1, \ldots, i_t), (j_1, \ldots, j_t)} \triangleq \prod_{t=1}^t A_{i_t, j_t}$ for every $(i_1, \ldots, i_t), (j_1, \ldots, j_t) \in \{1, \ldots, d\}^t$. We will use the following standard fact which in particular says that the spectral norm is multiplicative with respect to tensoring.

**Fact 5.6.13** \cite[e.g.,]{Lau05}. For any matrix $A \in \mathbb{R}^{d \times d}$, vector $u \in \mathbb{R}^d$, scalar $c \in \mathbb{R}$ and positive integer $t$, we have that

1. $\|cA\| = |c| \cdot \|A\|$. 

132
2. \( \|A^g\| = \|A\|^g \).

3. \( \|Au\|_2 \leq \|A\| \cdot \|u\|_2 \), where for any vector \( w \in \mathbb{R}^d \), \( \|w\|_2 \) denotes the Euclidean norm of \( w \), i.e., \( \|w\|_2 = \sqrt{\sum_{i=1}^{d} w_i^2} \).

The next lemma upper-bounds the spectral norm of an explicit family of \( 4 \times 4 \) matrices that will be used in the proof of Lemma 5.6.11. Looking ahead, it is crucial for our purposes that the coefficient multiplying \( a \) on the right-hand side of Equation (5.7) is a constant smaller than 2.

**Lemma 5.6.14.** Let \( a \in (0, 1) \) be a real number and \( N \triangleq N(a) \triangleq \begin{bmatrix} 1 & a & a & -a^2 \\ a & 1 & -a^2 & a \\ a & a^2 & 1 & -a \\ a^2 & a & -a & 1 \end{bmatrix} \).

Then,

\[ \|N\| \leq 1 + \sqrt{2} \cdot a + a^2 + \frac{a^4}{2} + \frac{a^5}{\sqrt{2}}. \quad (5.7) \]

The proof of Lemma 5.6.14 is deferred to the end of this section. We are now ready to prove Lemma 5.6.11.

**Proof of Lemma 5.6.11.** Fix any rectangle \( R = C \times D \) where \( C, D \subseteq \{0, 1\}^{2n} \). We wish to show that \( \text{Disc}_{\nu_p}(R, F) \leq 2^{-\gamma_p n} \). First, note that \( \text{Disc}_{\nu_p}(R, F) = |1_C M 1_D| \) where \( 1_C \) and \( 1_D \) are the 0/1 indicator vectors of \( C \) and \( D \) (respectively) and \( M \) is the \( 2^{2n} \times 2^{2n} \) real matrix defined by

\[ M((S,x),(T,y)) \triangleq \nu_p((S,T),(x,y)) \cdot (-1)^{\mathbb{X}_T(x \oplus y)} = \frac{1}{2^{2n}} (1 - p)^{2n} (-1)^{\mathbb{X}_T(x \oplus y)} \left( \frac{p}{1 - p} \right)^{|S \oplus T| + |x \oplus y|} \]

for every \( S, x, T, y \in \{0, 1\}^n \). Letting \( a \triangleq p/(1 - p) \), we can write

\[ M((S,x),(T,y)) = \frac{1}{2^{2n}} (1 - p)^{2n} (N \otimes n)((S,x),(T,y)) \]

with \( N = N(a) \) being the \( 4 \times 4 \) real matrix defined by

\[ N((S_1,x_1),(T_1,y_1)) \triangleq \langle -1 \rangle^{T_1(1 \oplus y_1)} d^{S_1 \oplus T_1}| + |x_1 \oplus y_1| \]  

\[ (5.8) \]

---

16. We here use the symbols \( S \) and \( T \) to denote both subsets of \( [n] \) and the corresponding 0/1 indicator vectors.

17. In Equation (5.8), \( T_1(x_1 \oplus y_1) \) denotes the product of the bit \( T_1 \) and the bit \( (x_1 \oplus y_1) \). Moreover, since \( (S_1 \oplus T_1) \) is a single bit, its Hamming weight \( |S_1 \oplus T_1| \) is the same as its bit-value, and similarly for \( (x_1 \oplus y_1) \).
for all \( S_1, x_1, T_1, y_1 \in \{0, 1\} \). Using the third property listed in Fact 5.6.13, we get that

\[
\text{Disc}_{\nu p}(R, F) = |1^T C M 1_D| \leq \|1_C\|_2 \cdot \|M\| \cdot \|1_D\|_2 \\
\leq \sqrt{2^{2n}} \cdot \|M\| \cdot \sqrt{2^{2n}} = 2^{2n} \cdot \|M\|.
\]  

(5.9)

We now use the first two properties listed in Fact 5.6.13 to relate \( \|M\| \) to \( \|N\| \) as follows:

\[
\|M\| = \|\frac{1}{2^{2n}} (1-p)^{2n} N^{\otimes n}\| = \frac{1}{2^{2n}} (1-p)^{2n} \|N\|^{n}.
\]  

(5.10)

Using Equation (5.8), we can check that

\[
N = N(a) = \begin{bmatrix}
1 & a & a & -a^2 \\
a & 1 & -a^2 & a \\
a & a^2 & 1 & -a \\
a^2 & a & -a & 1
\end{bmatrix}.
\]

Applying Lemma 5.6.14 with \( a = p/(1-p) \) and \( p \) sufficiently small (e.g., less than 1/10), we get

\[
\|N\| \leq 1 + \sqrt{2} \cdot \left( \frac{p}{1-p} \right) + O(p^2).
\]  

(5.11)

Combining Equation (5.9), Equation (5.10) and Equation (5.11) above, we conclude that

\[
\text{Disc}_{\nu p}(R, F) \leq (1-p)^{2n} \cdot (1 + \sqrt{2} \cdot \left( \frac{p}{1-p} \right) + O(p^2))^n \\
= \left[ (1-p) \cdot (1 + p \cdot (\sqrt{2} - 1) + O(p^2)) \right]^n \\
= \left[ 1 - p \cdot (2 - \sqrt{2}) + O(p^2) \right]^n \\
\leq 2^{-\gamma p^n}
\]

for some absolute constant \( \gamma > 0 \).

We conclude this section by proving Lemma 5.6.14.

**Proof of Lemma 5.6.14.** One can verify that

\[
N^T N = \begin{bmatrix}
(a^2 + 1)^2 & 2a(a^2 + 1) & 2a(1 - a^2) & 0 \\
2a(a^2 + 1) & (a^2 + 1)^2 & 0 & 2a(1 - a^2) \\
2a(1 - a^2) & 0 & (a^2 + 1)^2 & -2a(a^2 + 1) \\
0 & 2a(1 - a^2) & -2a(a^2 + 1) & (a^2 + 1)^2
\end{bmatrix}.
\]

134
Assuming that \( a \in (0, 1) \), one can also verify that \( N^T N \) has as eigenvectors

\[
\begin{align*}
\mathbf{v}_1 & \triangleq \begin{bmatrix}
\frac{\sqrt{2a^4+1}}{1-a^2} \\
\frac{a^2+1}{1-a^2} \\
1 \\
0
\end{bmatrix}, \\
\mathbf{v}_2 & \triangleq \begin{bmatrix}
\frac{a^2+1}{1-a^2} \\
\frac{\sqrt{2a^4+1}}{1-a^2} \\
0 \\
1
\end{bmatrix}
\end{align*}
\]

with eigenvalue \( \lambda_1(a) \triangleq 2a^2 + a^4 + 2a\sqrt{2(a^4 + 1)} + 1 \), and

\[
\begin{align*}
\mathbf{v}_3 & \triangleq \begin{bmatrix}
\frac{\sqrt{2(a^4+1)}}{a^2-1} \\
\frac{a^2+1}{1-a^2} \\
1 \\
0
\end{bmatrix}, \\
\mathbf{v}_4 & \triangleq \begin{bmatrix}
\frac{a^2+1}{a^2-1} \\
\frac{\sqrt{2(a^4+1)}}{a^2-1} \\
0 \\
1
\end{bmatrix}
\end{align*}
\]

with eigenvalue \( \lambda_2(a) \triangleq 2a^2 + a^4 - 2a\sqrt{2(a^4 + 1)} + 1 \).

Note that for any value of \( a \in (0, 1) \), the vectors \( v_1, v_2, v_3 \) and \( v_4 \) are linearly independent and each of the eigenvalues \( \lambda_1(a) \) and \( \lambda_2(a) \) has multiplicity 2. Moreover, we have that \( \lambda_1(a) \geq \lambda_2(a) \). Hence,

\[
\|N\| = \sqrt{\lambda_1(a)} = \sqrt{2a^2 + a^4 + 2a\sqrt{2(a^4 + 1)} + 1}.
\]

Applying twice the fact that \( \sqrt{1 + x} \leq 1 + x/2 \) for any \( x \geq -1 \), we get that

\[
\|N\| = \sqrt{1 + 2a^2 + a^4 + 2a\sqrt{2(1+a^4)}}
\leq 1 + a^2 + \frac{a^4}{2} + a\sqrt{2(1+a^4)}
\leq 1 + a^2 + \frac{a^4}{2} + a\sqrt{2(1+a^4)}
= 1 + a\sqrt{2} + a^2 + \frac{a^4}{2} + \frac{a^5}{\sqrt{2}}.
\]

\[\square\]

### 5.7 Improved Lower Bound on One-Way Public-Coin Uncertain Protocols

In this section, we prove Theorem 5.1.4. We start by giving a high-level overview before giving the proof.
5.7.1 Overview of Proof of Theorem 5.1.4

As will be explained in detail below, the proof of Theorem 5.1.4 is based on a construction that leads to the question described next regarding the communication complexity of a particular block-composed function. Namely, consider the following “majority composed with subset-parity with side information” setup. Alice is given a sequence of subsets \( S \triangleq (S^{(i)} \subseteq [n])_{i \in [k]} \) and a sequence of strings \( x \triangleq (x^{(i)} \in \{0, 1\}^n)_{i \in [k]} \), and Bob is given a sequence of subsets \( T \triangleq (T^{(i)} \subseteq [n])_{i \in [k]} \) and a sequence of strings \( y \triangleq (y^{(i)} \in \{0, 1\}^n)_{i \in [k]} \). We consider the following distribution \( \mu \) on \( ((S, x), (T, y)) \). Independently for each \( i \in [k] \), we sample \( ((S^{(i)}, x^{(i)}), (T^{(i)}, y^{(i)})) \) as follows: \( S^{(i)} \) is a uniform-random subset and \( T^{(i)} \) is an \( \epsilon \)-noisy\(^{18} \) copy of \( S^{(i)} \), and independently \( x^{(i)} \) is a uniform-random string and \( y^{(i)} \) is an \( \epsilon \)-noisy copy of \( x^{(i)} \). Here, \( \epsilon \) is a positive parameter that can depend on \( n \) and \( k \). Alice and Bob wish to compute the function \( \text{MAJ} \circ \text{SubsetParity}((S, x), (T, y)) \triangleq \text{Sign}(\sum_{i=1}^{k} (-1)^{(T^{(i)}, x^{(i)} \oplus y^{(i)})}) \) where \( T^{(i)} \) denotes both the subset and its 0/1 indicator vector, the inner product is over \( \mathbb{F}_2 \), and \( x^{(i)} \oplus y^{(i)} \) is the coordinate-wise XOR of \( x^{(i)} \) and \( y^{(i)} \). What is the communication complexity of computing \( \text{MAJ} \circ \text{SubsetParity} \) with high probability over the distribution \( \mu \)? Later in this section, we prove the following lower bound.

**Lemma 5.7.1.** Any one-way protocol computing \( \text{MAJ} \circ \text{SubsetParity} \) with high probability over the distribution \( \mu \) should communicate \( \Omega(k \cdot \epsilon \cdot n) \) bits.

We next explain how Theorem 5.1.4 leads to the setup of Lemma 5.7.1 and then give an overview of the proof of Lemma 5.7.1.

**Reduction to Lemma 5.7.1.** The proof of Theorem 5.1.4 builds on the lower-bound construction that we used to prove Theorem 5.1.3 in Section 5.6, which we briefly recall next. Let \( \mu \) be the distribution over pairs \( (x, y) \in \{0, 1\}^{2n} \) where \( x \) is uniform-random and \( y \) is an \( \epsilon \)-noisy copy of \( x \) with \( \epsilon = \sqrt{\delta/n} \). Then, the mutual information between \( x \) and \( y \) satisfies \( I \approx n \). For each \( S \subseteq [n] \), consider the function \( f_S(x, y) \triangleq \langle S, x \oplus y \rangle \) where the inner product is over \( \mathbb{F}_2 \), \( x \oplus y \) denotes the coordinate-wise XOR of \( x \) and \( y \), and \( S \) is used to denote both the subset and its 0/1 indicator vector. Moreover, consider the class \( \mathcal{F} \) of all pairs of functions \( (f_S, f_T) \) where \( |S \triangle T| \leq \sqrt{\delta n} \). It can be seen that for such \( S \) and \( T \), the distance between \( f_S \) and \( f_T \) under \( \mu \) is at most \( \delta \). If Alice and Bob both know \( S \), then Alice can send the single bit \( \langle S, x \rangle \) to Bob who can then output the correct answer \( \langle S, x \oplus y \rangle = \langle S, x \rangle \oplus \langle S, y \rangle \). This means that the certain communication is 1 bit. Using the well-known discrepancy method, in Section 5.6, we proved a lower bound of \( \Omega(\sqrt{n}) \) bits on the communication of the associated uncertain problem. Since in this case \( I \approx n \), this in fact lower-bounds the uncertain communication by \( \Omega(\sqrt{T}) \) bits. For this construction, this lower bound turns out to be tight up to a logarithmic factor.

To improve the lower-bound from \( \sqrt{T} \) to \( \sqrt{k} \cdot \sqrt{T} \) in the one-way setup, we consider the following “block-composed” framework. Let \( \{f_{S^{(i)}}(x^{(i)}, y^{(i)}): i \in [k]\} \) be \( k \) independent copies of

\(^{18}\)This means that the indicator vector of \( T^{(i)} \) is obtained by independently flipping each coordinate of the indicator vector of \( S^{(i)} \) with probability \( \epsilon \).
of the above base problem of [GKKS16] and consider computing the composed function 
\( g(f_{S(1)}(x^{(1)}, y^{(1)}), \ldots, f_{S(k)}(x^{(k)}, y^{(k)})) \) for some outer function \( g : \{0, 1\}^k \to \{0, 1\} \). For any choice of \( g \), the certain communication of the composed function would be at most \( k \) bits. When choosing the outer function \( g \) to use in our lower bound, we thus have two objectives to satisfy. First, \( g \) has to be sufficiently hard in the sense that its average-case decision tree complexity with respect to the uniform distribution on \( \{0, 1\}^k \) should be \( \Omega(k) \); otherwise, it will not be the case that the uncertain communication of computing \( g \) on \( k \) copies of the base problem is at least \( k \) times the uncertain communication of the base problem. Second, \( g \) has to be noise stable in order to be able to upper-bound the distance between 
\( g(f_{S(1)}(\cdot), \ldots, f_{S(k)}(\cdot)) \) and 
\( g(f_{T(1)}(\cdot), \ldots, f_{T(k)}(\cdot)) \).

Note that setting \( g \) to be a dictator function would satisfy the noise-stability property, but it clearly would not satisfy the hardness property, as the composed function would be equal to the base function and would thus have uncertain communication \( \tilde{O}(\sqrt{n}) \) bits. Another potential choice of \( g \) is to set it to the parity function on \( k \) bits. This function would satisfy the hardness property, but it would strongly violate the noise stability property that is crucial to us. This leads us to setting \( g \) to the majority function on \( k \) bits, which is well-known to be noise stable, and has average-case decision-tree complexity \( \Omega(k) \) with respect to the uniform distribution on \( \{0, 1\}^k \). In fact, the noise stability of the majority function readily implies an upper bound of \( O(\sqrt{k}) \) on the distance between any pair of composed functions that are specified by tuples of subsets \( (S^{(1)}, \ldots, S^{(k)}) \) and \( (T^{(1)}, \ldots, T^{(k)}) \) with \( |S^{(i)} \Delta T^{(i)}| \leq \sqrt{kn} \) for each \( i \in [k] \). The crux of the proof will be to lower-bound the uncertain communication of the majority-composed function by \( \Omega(k\sqrt{n}) \), which amounts to proving Lemma 5.7.1. Since in this block-composed framework the mutual information satisfies \( I \approx kn \), this would imply the lower bound of \( \Omega(\sqrt{k} \sqrt{T}) \) on the uncertain communication in Part (iii) of Theorem 5.1.4.

Overview of Proof of Lemma 5.7.1. We first point out that the average-case quantum decision tree complexity of \( \text{MAJ}_k \) with respect to the uniform distribution is \( \tilde{O}(\sqrt{k}) \) [ADW01]. This implies that any communication complexity lower bound method that extends to the quantum model cannot prove a lower bound larger than \( \tilde{O}(\sqrt{k} \cdot \sqrt{n}) \) on our uncertain communication\(^{19}\). In particular, we cannot solely rely on the discrepancy bound (as we did in the proof of Theorem 5.1.4), since this bound is known to lower-bound quantum communication. Similarly, the techniques of [She08, SZ07, LZ10] rely on the generalized discrepancy bound (originally due to [Kla01]) which also lower-bounds quantum communication. Moreover, the recent results of [MWY15] only apply to product distributions (i.e., where Alice’s input is independent of Bob’s input) in contrast to our case where the inputs of Alice and Bob are very highly-correlated. Finally, the recent works of [GLM+15, GPW15] do not imply lower bounds on the average-case complexity with respect to the distribution that arises in our setup.

To circumvent the above obstacles, we use a new approach that is tailored to our setup and that we outline next. Let \( \Pi \) be a one-way protocol solving the uncertain task with

\(^{19}\text{Thus, since } I \approx k \cdot n \text{ in our block-composed framework, such methods cannot be used to improve the lower-bound of } \Omega(\sqrt{T}) \text{ in Theorem 5.1.4 by more than logarithmic factors.} \)
high probability. We consider the information that \( \Pi \) reveals about the inputs to the outer function, i.e., about the length-\( k \) binary string \( (f_{S(i)}(x^{(1)}, y^{(1)}), \ldots, f_{S(k)}(x^{(k)}, y^{(k)})) \). We call this quantity the intermediate information cost of \( \Pi \), and we argue that it is at least \( \Omega(k) \) bits. To do so, we recall the Hamming distance function \( \text{HD}_k \) defined by \( \text{HD}_k(u, v) = 1 \) if the Hamming distance between \( u \) and \( v \) is at least \( k/2 \) and \( \text{HD}_k(u, v) = 0 \) otherwise. We upper-bound the information complexity of computing \( \text{HD}_k \) over the uniform distribution on \( \{0, 1\}^{2k} \) by the intermediate information cost of \( \Pi \). We do so by giving an information-cost preserving procedure (Protocol 4) where Alice and Bob are given independent uniformly distributed \( u \) and \( v \) (respectively) and use their private and public coins in order to simulate the input distribution \( (X, Y) \) of our uncertain problem. The known one-way lower bound of [Woo07] on \( \text{HD}_k \) under the uniform distribution then implies that \( \Pi \) reveals \( \Omega(k) \) bits of information to Bob about the tuple \( (f_{S(i)}(x^{(1)}, y^{(1)}), \ldots, f_{S(k)}(x^{(k)}, y^{(k)})) \). This allows Bob to guess this tuple with probability \( 0.51^k \). We then apply the strong direct product theorem for discrepancy of [LSS08] which, along with the discrepancy-based lower bound on the communication of the base uncertain problem given in Theorem 5.1.4, implies that \( \Pi \) should be communicating at least \( \Omega(k \sqrt{n}) \) bits.

We now turn to the formal proof of Theorem 5.1.4.

5.7.2 Proof of Theorem 5.1.4

We start by formally describing the construction that is used to prove Theorem 5.1.4. We set

\[ \delta' \triangleq c \cdot \delta^2 \quad \text{and} \quad \epsilon \triangleq \sqrt{\delta'/n}, \tag{†} \]

where \( \delta \) is the parameter from the statement of Theorem 5.1.4, and \( c > 0 \) is a small-enough absolute constant. To define our input distribution, we first define a slightly more general distribution \( \mu_\eta \). The support of \( \mu_\eta \) is \( \{0, 1\}^{kn} \times \{0, 1\}^{kn} \) and we will view the coordinates of a sample \( (x, y) \sim \mu_\eta \) as \( x = (x^{(i)})_{i \in [k]} \) and \( y = (y^{(i)})_{i \in [k]} \) with \( x^{(i)}, y^{(i)} \in \{0, 1\}^n \) for all \( i \in [k] \). A sample \( (x, y) \sim \mu_\eta \) is generated by letting \( x \in \mathcal{R} \{0, 1\}^{kn} \) and for all \( i \in [k] \) and \( j \in [n] \), independently setting \( y^{(i)}_j \) to be an \( \eta \)-noisy copy of \( x^{(i)}_j \). In other words, we set \( y^{(i)}_j = \hat{x}^{(i)}_j \) with probability \( 1 - \eta \) and \( y^{(i)}_j = 1 - \hat{x}^{(i)}_j \) with probability \( \eta \). Our input distribution is then \( \mu_{2\epsilon - 2\epsilon^2} \).

We now define our function class \( \mathcal{F}_\epsilon \). Each function in our universe is specified by a sequence of subsets \( S \triangleq (S^{(i)} \subseteq [n])_{i \in [k]} \) and it is of the form \( f_S : \{0, 1\}^{2kn} \to \{0, 1\} \) with\(^{20}\)

\[ f_S(x, y) \triangleq \text{Sign}\left( \sum_{i \in [k]} (-1)^{\langle S^{(i)}, x^{(i)} \oplus y^{(i)} \rangle} \right) \tag{5.12} \]

for all \( x, y \in \{0, 1\}^{kn} \), where in Eq. (5.12) the inner product is over \( \mathbb{F}_2 \), the sum is over \( \mathbb{R} \) and \( x^{(i)} \oplus y^{(i)} \) denotes the coordinate-wise XOR of the two length-\( n \) binary strings \( x^{(i)} \) and \( y^{(i)} \). The function class is then defined by \( \mathcal{F}_\epsilon \triangleq \{(f_S, f_T) : |S^{(i)} \triangleq T^{(i)}| \leq \epsilon \cdot n \text{ for all } i \in [k]\}.\]

\(^{20}\)We will use the symbol \( S^{(i)} \) to denote both the subset of \( [n] \) and its corresponding 0/1 indicator vector.
We now give the proof of Part (i) of Theorem 5.1.4. It follows from known bounds on the noise stability of the majority function.

*Proof of Part (i) of Theorem 5.1.4.* Let \((f_S, f_T) \in \mathcal{F}_e\), and denote \(a_i \triangleq (-1)^{(S(i), x(i) \oplus y(i))}\) and \(b_i \triangleq (-1)^{(T(i), x(i) \oplus y(i))}\) for every \(i \in [k]\). Also, let \(a \triangleq (a_i)_{i \in [k]}\) and \(b \triangleq (b_i)_{i \in [k]}\). Note that \((a_i, b_i)\) is a pair of \(\rho_i\)-correlated random strings with \(\rho_i \geq (1 - 2\delta')\). Since \(\delta_{\mu_{2\epsilon-2\epsilon^2}}(f_S, f_T)\) increases when \(\rho_i\) decreases, we assume without loss of generality that \(\rho_i = 1 - 2\delta' \triangleq \rho\) for all \(i \in [k]\).

Recall that the noise stability of a function \(h : \{0, 1\}^k \to \{\pm 1\}\) is defined as \(\text{Stab}_\rho(h) = \mathbb{E}[h(x)h(y)]\) where \((x, y)\) is a random pair of \(\rho\)-correlated strings. Let the function \(\widetilde{\text{MAJ}}_k\) be defined by \(\widetilde{\text{MAJ}}_k(x) = (-1)^{\text{MAJ}_k(x)}\) for all \(x \in \{0, 1\}^k\). Recall (see [O'D14]) that the noise stability of \(\widetilde{\text{MAJ}}_k\) satisfies

\[
\text{Stab}_\rho(\widetilde{\text{MAJ}}_k) \geq 1 - \frac{2}{\pi} \arccos(\rho).
\]

Hence, we get that

\[
\delta_{\mu_{2\epsilon-2\epsilon^2}}(f_S, f_T) = \Pr[f_S(x, y) \neq f_T(x, y)] \\
\leq \frac{1 - \text{Stab}_\rho(\widetilde{\text{MAJ}}_k)}{2} \\
\leq \frac{\arccos(\rho)}{\pi} \\
= O(\sqrt{\delta'}) \\
\leq \delta,
\]

where the last equality uses the facts that \(\rho = 1 - 2\delta'\) and that \(\arccos(1 - x) = O(\sqrt{x})\) for small positive values of \(x\), and the last inequality follows from the setting of \(\delta'\) in (†).

We now give the (straightforward) proof of Part (ii) of Theorem 5.1.4.

*Proof of Part (ii) of Theorem 5.1.4.* Note that \(\langle S(i), x(i) \oplus y(i) \rangle = \langle S(i), x(i) \rangle \oplus \langle S(i), y(i) \rangle\). Thus, when both Alice and Bob know \(S\), Alice can send the \(k\) bits \((\langle S(i), x(i) \rangle)_{i \in [k]}\) to Bob who can then output the value \(f_S(x, y)\).

In order to prove Part (iii) of Theorem 5.1.4, we first define (as in Section 5.6) a communication problem in the standard distributional model that reduces to solving the contextually-uncertain problem specified by the function class \(\mathcal{F}_e\) and the distribution \(\mu_{2\epsilon-2\epsilon^2}\). For distributions \(\phi\) and \(\psi\), we denote by \(\phi \otimes \psi\) the joint distribution of a sample from \(\phi\) and an independent sample from \(\psi\). The new problem is defined as follows.

**Inputs:** Alice’s input is a pair \((S, x)\) where \(S \triangleq (S(i) \subseteq [n])_{i \in [k]}\) and \(x \in \{0, 1\}^kn\). Bob’s input is a pair \((T, y)\) where \(T \triangleq (T(i) \subseteq [n])_{i \in [k]}\) and \(y \in \{0, 1\}^kn\).
**Distribution:** Let $D_q$ be the distribution on the pair $(S, T)$ of sequences of $k$ subsets of $[n]$, which is defined by independently setting, for each $i \in [k]$, $S^{(i)}$ to be a uniformly-random subset of $[n]$, and $T^{(i)}$ to be a $q$-noisy copy of $S^{(i)}$. The distribution on the inputs of Alice and Bob is then given by $\nu_\epsilon \triangleq D_q \otimes \mu_{2k-2\epsilon^2}$ with $\epsilon = \sqrt{\delta/n}$.

**Function:** The goal is to compute the function $F : \{0, 1\}^{2kn} \times \{0, 1\}^{2kn} \to \{0, 1\}$ defined by $F((S, x), (T, y)) = \text{Sign}( \sum_{i \in [k]} (-1)^{f_T(x, S^{(i)}; y^{(i)})} )$.

The next proposition follows from a simple application of the Chernoff bound.

**Proposition 5.7.2.** For any $\theta > 0$, $\text{owPubCC}_{\theta}^{2kn-2-\epsilon^2}(F) \geq \text{owCC}_{\theta+\epsilon'}^{\nu_k}(F)$ with $\epsilon' = 2^{-\Theta(\epsilon n)}$.

We will prove the following lower bound on $\text{owCC}_{\theta}^{\nu_k}(F)$, which along with Proposition 5.7.2 and the settings of $\epsilon$ and $\delta'$ in $(\dagger)$, implies Part (iii) of Theorem 5.1.4:

**Lemma 5.7.3.** For every sufficiently small positive constant $\theta$, $\text{owCC}_{\theta}^{\nu_k}(F) = \Omega(k \cdot \epsilon \cdot n)$.

We now prove Lemma 5.7.3 (which is the bulk of the proof of Theorem 5.1.4). Section 5.7.3 summarizes some known results that we use in Sections 5.7.4 and 5.7.5. In Section 5.7.4, we prove a “Simulation Lemma” that will be useful to us. In Section 5.7.5, we prove Lemma 5.7.3.

### 5.7.3 Proof Preliminaries for Sections 5.7.4 and 5.7.5

In this section, we state some tools and known results that we use in the proofs in Sections 5.7.4 and 5.7.5. We use the following strong direct product theorem for discrepancy of [LSS08].

**Lemma 5.7.4** (Corollary 23 of [LSS08]). Let $f : X \times Y \to \{0, 1\}$ be a Boolean function and $P$ a probability distribution over $X \times Y$. If $\text{CC}_{1/2-w/2}(f) \geq C$ is proved using the discrepancy method, then the success probability under distribution $P^{\otimes k}$ of any $kC/3$ bit protocol computing the vector of solutions $f^{(k)}$ is at most $(8w)\tau-k + 2^{-k(1-H_0(\tau))}$ where $\tau$ is any positive constant less than 0.5.

Let $\xi_\epsilon$ be the distribution that is obtained by projecting $\nu_\epsilon$ on one of the $k$ blocks and marginalizing over the remaining $k-1$ blocks. Namely, $\nu_\epsilon = \xi_\epsilon^{\otimes k}$. Define the “base function” $G((\bar{S}, \bar{x}), (\bar{T}, \bar{y}))$ by $G((\bar{S}, \bar{x}), (\bar{T}, \bar{y})) = (\bar{T}, \bar{x} \oplus \bar{y})$ for every $\bar{S}, \bar{T} \subseteq [n]$ and $\bar{x}, \bar{y} \in \{0, 1\}^n$. We will use the following lower bound on the distributional communication complexity of $G$ over $\xi_\epsilon$ which is a direct corollary of Theorem 5.1.3 and which was proved using the discrepancy method.

\[^{21}\text{We point out that the statement of Corollary 23 of [LSS08] has a small inaccuracy: the additive } 2^{-k(1-H_0(\tau))} \text{ term in Lemma 5.7.4 is missing. This term is clearly needed as one can always guess } f^{(k)} \text{ with probability } 2^{-k}. \]
Lemma 5.7.5 (Corollary of Theorem 5.1.3). For any \( w = 2^{-w(\sqrt{\frac{\Omega}{n}})} \), we have that \( CC_{\frac{1}{2} - w/2}(G) \geq \Omega(\varepsilon \cdot n) \), and it is proved using the discrepancy method.

Combining Lemma 5.7.4 and Lemma 5.7.5 implies the next corollary.

Corollary 5.7.6. For every positive constant \( \gamma \), any deterministic protocol computing \( G(k) \) correctly with probability at least \( (0.5 + \gamma)^k \) with respect to the distribution \( \xi^\varepsilon_k = \nu_e \) should be communicating \( \Omega(k \cdot \varepsilon \cdot n) \) bits.

We define the Hamming distance function \( HD_k : \{0, 1\}^k \times \{0, 1\}^k \rightarrow \{0, 1\} \) as follows. For all \( x, y \in \{0, 1\}^k \), \( HD_k(x, y) = 1 \) if the Hamming distance between \( x \) and \( y \) is at least \( \lceil k/2 \rceil \) and \( HD_k(x, y) = 0 \) otherwise. Let \( \mathcal{U}_{2k} \) denote the uniform distribution on \( \{0, 1\}^{2k} \).

Lemma 5.7.7 ([Woo07]). For every sufficiently small \( \varepsilon > 0 \), it holds that \( owCC_{\varepsilon}^{\mathcal{U}_{2k}}(HD_k) = \Omega(k) \).

The next lemma of [JRS03] compresses a one-way private-coin protocol with external information cost\(^{22} I \) into a one-way deterministic protocol with communication cost \( O(I) \).

Lemma 5.7.8 (Result 1 of [JRS03]). Suppose that \( \Pi \) is a one-way private-coin randomized protocol for \( f : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{Z} \). Let the average error of \( \Pi \) under a probability distribution \( \mu \) on the inputs \( \mathcal{X} \times \mathcal{Y} \) be \( \theta \). Let \( X, Y \) denote the random variables corresponding to Alice’s and Bob’s inputs respectively. Let \( M \) denote the single message sent by Alice to Bob. Suppose \( I(X, Y; M) \leq a \). Let \( \zeta > 0 \). Then, there is another deterministic one-way protocol \( \Pi' \) with the following properties:

1. The communication cost of \( \Pi' \) is at most \( \frac{2(\alpha + 1)}{\zeta} + \frac{2}{\zeta} \) bits.
2. The distributional error of \( \Pi' \) under \( \mu \) is at most \( \theta + 2\zeta \).

We will also use the next lemma.

Lemma 5.7.9. Let \( (Q, W, B) \) be correlated random variables with \( Q \in \mathcal{Q}, W \in \mathcal{W} \) and \( B \in \{0, 1\}^k \). Let \( \alpha \in (0, 1] \) be any constant. If \( I(B; W|Q) \geq \alpha \cdot k \), then there exists a positive constant \( \beta \) that only depends on \( \alpha \), and a deterministic function \( E : \mathcal{Q} \times \mathcal{W} \rightarrow \{0, 1\}^k \) such that \( E(Q, W) = B \) with probability at least \( (0.5 + \beta)^k \) over the random choice of \( (Q, W, B) \).

Proof of Lemma 5.7.9. Consider the deterministic function \( E : \mathcal{Q} \times \mathcal{W} \rightarrow \{0, 1\}^k \) defined as follows. For each \( (q, w) \in \mathcal{Q} \times \mathcal{W} \), \( E(q, w) \) is set to an arbitrary element of the set \( \arg \max_{B \in \{0, 1\}^k} \Pr[B = \hat{B}|Q = q, W = w] \).

\(^{22}\)The external information cost of a protocol is the amount of information that it reveals about the inputs to an external observer. For a one-way private-coin protocol, it is given by \( I(X, Y; M) \) where \( M \) is the single message sent from Alice to Bob.
We now argue that $E(Q, W) = B$ with probability at least $(0.5 + \beta)^k$ over the randomness of $(Q, W, B)$, where $\beta$ is a positive constant that only depends on $\alpha$. Since $I(B; W|Q) \geq \alpha \cdot k$, we have that

$$H(B|Q, W) = H(B|Q) - I(B; W|Q)$$

$$\leq H(B|Q) - \alpha \cdot k$$

$$\leq H(B) - \alpha \cdot k$$

$$\leq (1 - \alpha) \cdot k,$$

where the third inequality above uses the fact that conditioning does not increase entropy, and the fourth inequality follows from the fact that $B \in \{0, 1\}^k$. By an averaging argument, with probability at least $\alpha/10$ over $(q, w) \sim (Q, W)$, it should be the case that

$$H(B|Q = q, W = w) \leq (1 - \alpha/10) \cdot k.$$  \hspace{1cm} (5.13)

Let $\mathcal{G} \subseteq Q \times W$ denote the set of all pairs $(q, w)$ that satisfy Equation (5.13). We now fix $(q, w) \in \mathcal{G}$, and consider the min-entropy

$$H_{\min}(B|Q = q, W = w) \triangleq \min_{B \in \{0, 1\}^k} \log_2 \left( \frac{1}{\Pr[B = B|Q = q, W = w]} \right).$$

Using the fact that min-entropy lower-bounds Shannon entropy and Equation (5.13), we deduce that there exists $\bar{B} \triangleq \bar{B}(q, w) \in \{0, 1\}^k$ such that $\Pr[B = \bar{B}|Q = q, W = w] \geq 2^{-(1-\alpha/10)^k}$. Hence, for any fixed $(q, w) \in \mathcal{G}$, conditioned on $(Q = q, W = w)$, the value $E(Q, W)$ is equal to $B$ with probability at least $2^{-(1-\alpha/10)^k}$.

Since the probability that $(Q, W) \in \mathcal{G}$ is at least $\alpha/10$, we conclude that $E(Q, W)$ is equal to $B$ with probability at least

$$(\alpha/10) \cdot 2^{-(1-\alpha/10)^k} \geq (0.5 + \beta)^k,$$

for some constant $\beta$ that only depends on $\alpha$. \hfill \Box

### 5.7.4 Simulation Protocol

Recall that the distribution $\nu_\epsilon$ over the inputs $((S, X), (T, Y))$ to $F$ was defined as $\nu_\epsilon \triangleq \mathcal{D}_\epsilon \otimes \mu_{2\epsilon - 2\epsilon^2}$. In the following simulation lemma, the error probability will be measured with respect to distribution $\nu_\epsilon$ whereas the information cost will be measured with respect to another distribution $\kappa_\epsilon$ over $((S, X), (T, Y))$ inputs, which is defined as $\kappa_\epsilon \triangleq \mathcal{D}_\epsilon \otimes \mu_\epsilon$.

**Lemma 5.7.10** (Simulation Lemma). Let $\Pi$ be any deterministic one-way protocol computing $F$ with error at most $\theta$ on the distribution $\nu_\epsilon$ over $((S, X), (T, Y))$ inputs, and let $M \triangleq M(X, S)$ be the corresponding single message that is sent from Alice to Bob under $\Pi$.

Then, we have that

$$I_{((S, X), (T, Y)) \sim \nu_\epsilon} \left( ((T^{(i)}), X^{(i)}); i \in [k]; M(X, S) \mid Y, T \right) \geq \beta \cdot k$$  \hspace{1cm} (5.14)
for some constant $\beta > 0$ that only depends on $\theta$.

We point out that in Lemma 5.7.10, the error probability is measured with respect to the distribution $\nu$, while the information cost is measured with respect to the distribution $\kappa$.

**Definition 5.7.11.** A sequence $\hat{T} \triangleq (\hat{T}^{(i)} \subseteq [n])_{i \in [k]}$ of subsets is said to be typical if $|\hat{T}^{(i)}| \in [n/3, 2n/3]$ for all $i \in [k]$.

Recall that the total variation distance between two distributions $\phi$ and $\psi$ defined on the same finite support $\Omega$ is given by $\Delta_{TV}(\phi, \psi) = \max_{A \subseteq \Omega} |\phi(A) - \psi(A)| = 0.5 \cdot \sum_{x \in \Omega} |\phi(x) - \psi(x)|$. We will use the next lemma.

**Lemma 5.7.12** (Closeness Lemma). For a given sequence $\hat{T} \triangleq (\hat{T}^{(i)} \subseteq [n])_{i \in [k]}$ of subsets, we define the distribution $\mu_{\hat{T}, \epsilon}$ as follows. To sample $\langle X, Y \rangle \sim \mu_{\hat{T}, \epsilon}$, we independently sample $U, V \in_R \{0, 1\}^k$, $Z \in_R \{0, 1\}^{k-n}$, $X$ to be an $\epsilon$-noisy copy of $Z$ conditioned on $\langle \hat{T}^{(i)}, X^{(i)} \rangle_{i \in [k]} = U$, and $Y$ to be an $\epsilon$-noisy copy of $Z$ conditioned on $\langle \hat{T}^{(i)}, Y^{(i)} \rangle_{i \in [k]} = V$.

Then, for every fixed typical $\hat{T}$, we have that

$$\Delta_{TV}(\mu_{2\epsilon-2\epsilon^2}, \mu_{\hat{T}, \epsilon}) \leq k \cdot \exp(-\epsilon \cdot n).$$

**Proof of Lemma 5.7.12.** We denote $\langle \hat{T}, X \rangle \triangleq (\langle \hat{T}^{(i)}, X^{(i)} \rangle)_{i \in [k]}$ and similarly $\langle \hat{T}, Y \rangle \triangleq (\langle \hat{T}^{(i)}, Y^{(i)} \rangle)_{i \in [k]}$. Note that both $\langle \hat{T}, X \rangle$ and $\langle \hat{T}, Y \rangle$ are elements of $\{0, 1\}^k$. For every fixed $\hat{X}, \hat{Y} \in \{0, 1\}^{k-n}$, define $\hat{U} \triangleq \langle \hat{T}, \hat{X} \rangle$ and $\hat{V} \triangleq \langle \hat{T}, \hat{Y} \rangle$. We have that

$$\mu_{2\epsilon-2\epsilon^2}(X = \hat{X}, Y = \hat{Y}) = \frac{1}{2^{k-n}} \cdot (2\epsilon - 2\epsilon^2)^\Delta(\hat{X}, \hat{Y}) \cdot (1 - 2\epsilon + 2\epsilon^2)^{k-n-\Delta(\hat{X}, \hat{Y})}.$$

On the other hand, we have that

$$\mu_{\hat{T}, \epsilon}(X = \hat{X}, Y = \hat{Y}) = \mu_{\hat{T}, \epsilon}(U = \hat{U}, V = \hat{V}, X = \hat{X}, Y = \hat{Y}) = \sum_{\hat{Z} \in \{0, 1\}^{k-n}} \mu_{\hat{T}, \epsilon}(Z = \hat{Z}, U = \hat{U}, V = \hat{V}, X = \hat{X}, Y = \hat{Y})$$

$$= \sum_{\hat{Z} \in \{0, 1\}^{k-n}} \mu_{\hat{T}, \epsilon}(Z = \hat{Z}, U = \hat{U}, V = \hat{V}) \cdot \mu_{\hat{T}, \epsilon}(X = \hat{X}, Y = \hat{Y} | Z = \hat{Z}, U = \hat{U}, V = \hat{V})$$

$$= \sum_{\hat{Z} \in \{0, 1\}^{k-n}} \frac{1}{4^k} \cdot \frac{1}{2^{k-n}} \cdot \mu_{\hat{T}, \epsilon}(X = \hat{X} | Z = \hat{Z}, U = \hat{U}) \cdot \mu_{\hat{T}, \epsilon}(Y = \hat{Y} | Z = \hat{Z}, V = \hat{V}) \quad (5.15)$$

Denote by $N_\epsilon(Z)$ the distribution of a random variable that is an $\epsilon$-noisy copy of $Z$. Then,

$$\mu_{\hat{T}, \epsilon}(X = \hat{X} | Z = \hat{Z}, U = \hat{U}) = \frac{\Pr_{X' \sim N_\epsilon(Z)}[X' = \hat{X}]}{\Pr_{X' \sim N_\epsilon(Z)}[\langle \hat{T}, X' \rangle = \hat{U}]} \quad (5.16)$$

143
where

\[ \Pr_{X' \sim N_i(Z)} [(\hat{T}, X') = \hat{U}] = \prod_{i=1}^{k} \Pr_{X' \sim N_i(Z)} [(\hat{T}^{(i)}, X'^{(i)}) = \hat{U}^{(i)}] \]

\[ = \prod_{i=1}^{k} \frac{1}{2} \pm \exp(-\epsilon \cdot n) \]

\[ = \frac{1}{2^k} \cdot \prod_{i=1}^{k} (1 \pm \exp(-\epsilon \cdot n)) \]

\[ = \frac{1}{2^k} \cdot (1 \pm k \cdot \exp(-\epsilon \cdot n)), \]

where the third equality above follows from the fact that \( \hat{T} \) is typical. Plugging back this last expression in Equation (5.16), we get

\[ \mu_{T,\epsilon}^*(X = \hat{X}, Y = \hat{Y}) = 2^k \cdot \Pr_{X' \sim N_i(Z)} [X' = \hat{X}] \cdot (1 \pm k \cdot \exp(-\epsilon \cdot n)). \]

(5.17)

Similarly, we have that

\[ \mu_{T,\epsilon}^*(Y = \hat{Y}|Z = \hat{Z}, V = \hat{V}) = 2^k \cdot \Pr_{Y' \sim N_i(Z)} [Y' = \hat{Y}] \cdot (1 \pm k \cdot \exp(-\epsilon \cdot n)). \]

(5.18)

Combining Equations (5.15), (5.17) and (5.18) yields

\[ \mu_{T,\epsilon}^*(X = \hat{X}, Y = \hat{Y}) = \sum_{\hat{Z} \in \{0,1\}^{k \cdot n}} \frac{1}{2^{k \cdot n}} \cdot \Pr_{X' \sim N_i(Z)} [X' = \hat{X}] \cdot \Pr_{Y' \sim N_i(Z)} [Y' = \hat{Y}] \cdot (1 \pm k \cdot \exp(-\epsilon \cdot n))^2 \]

\[ = (1 \pm k \cdot \exp(-\epsilon \cdot n)) \cdot \sum_{\hat{Z} \in \{0,1\}^{k \cdot n}} \frac{1}{2^{k \cdot n}} \cdot \Pr_{X' \sim N_i(Z)} [X' = \hat{X}] \cdot \Pr_{Y' \sim N_i(Z)} [Y' = \hat{Y}] \]

\[ = (1 \pm k \cdot \exp(-\epsilon \cdot n)) \cdot \mu_{2\epsilon - 2\epsilon^2}(X = \hat{X}, Y = \hat{Y}), \]

(5.19)

where the last equality above follows from the fact that one way to sample a \((2\epsilon - 2\epsilon^2)\)-noisy pair \((X, Y)\) is to first sample a uniform-random \(Z\), and then independently sample each of \(X\) and \(Y\) to be an \(\epsilon\)-noisy copy of \(Z\). Using Equation (5.19) and the definition of the total
variation distance, we conclude that

\[
\Delta_{TV}(\mu_{2^{2\epsilon}}, \mu_{\hat{T}, \epsilon}) = \frac{1}{2} \cdot \sum_{\hat{X}, \hat{Y} \in \{0, 1\}^{k \cdot n}} |\mu_{2^{2\epsilon}}(X = \hat{X}, Y = \hat{Y}) - \mu_{\hat{T}, \epsilon}(X = \hat{X}, Y = \hat{Y})| \\
\leq k \cdot \exp(-\epsilon \cdot n) \sum_{\hat{X}, \hat{Y} \in \{0, 1\}^{k \cdot n}} \mu_{2^{2\epsilon}}(X = \hat{X}, Y = \hat{Y}) \\
= k \cdot \exp(-\epsilon \cdot n). \tag{5.20}
\]

We are now ready to prove Lemma 5.7.10.

Proof of Lemma 5.7.10. Assume for the sake of contradiction that there exists a deterministic one-way protocol \( \Pi \) that computes \( F \) with error at most \( \theta \) on \( \nu_{\epsilon} \) and that violates Equation (5.14). Namely, if we define the intermediate information cost of \( \Pi \) as

\[
\text{IIC}_{\nu_{\epsilon}}(\Pi | T) \triangleq I_{(S, X, (T, Y)) \sim \nu_{\epsilon}} \left( (\langle T^{(i)}, X^{(i)} \rangle)_{i \in [k]}; M(X, S) | Y, T \right),
\]

then we assume that \( \text{IIC}_{\nu_{\epsilon}}(\Pi | T) = o(k) \). We say that a particular value \( \hat{T} \) of \( T \) is nice if it simultaneously satisfies the following three properties:

1. \( \hat{T} \) is typical.
2. The conditional error of \( \Pi \) with respect to \( \nu_{\epsilon} \) conditioned on \( T = \hat{T} \) is at most \( O(\theta) \).
3. The intermediate information cost conditioned on \( T = \hat{T} \) satisfies

\[
\text{IIC}_{\nu_{\epsilon}}(\Pi | T = \hat{T}) = O\left( \text{IIC}_{\nu_{\epsilon}}(\Pi | T) \right) = o(k).
\]

We now argue that there exists a \( \hat{T} \) that is nice. To do so, we show that a random \( \hat{T} \) satisfies the above three properties with high probability. First, by Definition 5.7.11, a Chernoff bound and union bound, a random \( \hat{T} \) satisfies property 1 with probability at least \( 1 - o(1) \) as long as \( k \cdot \exp(-n) = o(1) \). Moreover, by an averaging argument, a random \( \hat{T} \) satisfies property 2 with probability \( 1 - o(1) \). Finally, by an averaging argument and the definition of the conditional mutual information in Equation (5.20), we get that a random \( \hat{T} \) satisfies property 3 with probability \( 1 - o(1) \). By a union bound, we conclude that a random \( \hat{T} \) satisfies all three properties with high probability. Henceforth, we fix such a nice \( \hat{T} \) and use it to give a deterministic one-way protocol computing the function \( \mathsf{HD}_{k} \) w.h.p. over the uniform distribution on \( \{0, 1\}^{2^{k}} \) and with communication \( o(k) \) bits. This would contradict the lower bound of [Woo07] (i.e., Lemma 5.7.7).

Consider the simulation protocol \( \Pi' \) described in Protocol 4. In this protocol, Alice is given as input a binary string \( U \) of length \( k \) and Bob is given as input a binary string \( V \) of length \( k \). We will argue that
(a) The output of $Π'$ is equal to $\text{HD}_k(U, V)$ with probability $1 - O(θ)$ over the randomness of $(U, V) \sim U_{2k}$ and over the private and shared randomness of $Π'$. 

(b) The information cost of $Π'$ satisfies

$$I(U; M'(U) \mid V, R) = \text{IIC}_{κ_4}(Π \mid T = ̂T),$$

where $M'$ is the single (randomized) message sent from Alice to Bob under $Π'$, and $R$ is the public randomness of $Π'$.

We start by proving property (a). Let $λ$ be the probability distribution of the sequence $S$ of subsets that is sampled in Protocol 4. In other words, $S \sim λ$ is an $ε$-noisy copy of $̂T$. Then, when $(U, V)$ is drawn uniformly at random, the induced distribution on $(S, X, Y)$ in Protocol 4 is $λ \otimes μ_{̂T, k}$. Property 2 above guaranteed that the error probability of protocol $Π$ on pairs $((S, X), (T, Y))$ such that $(S, X, Y) \sim λ \otimes μ_ε$ is at most $O(θ)$. Using Lemma 5.7.12, the fact that Protocol $Π'$ simulates $Π$ and the fact that

$$F((S, X), (̂T, Y)) = \text{HD}_k\left(\{(̂T^{(i)}, X^{(i)})\}_{i \in [k]}, \{(̂T^{(i)}, Y^{(i)})\}_{i \in [k]}\right),$$

we get that the error probability of $Π'$ (over the randomness of $(U, V) \sim U_{2k}$ and over the private and shared randomness) is at most $O(θ) + O(k \cdot \exp(-εn))$, which is $O(θ)$.

We next prove property (b). The information cost of $Π'$ is given by

$$I(U; M'(U) \mid V, R) = I((̂T^{(i)}, X^{(i)})_{i \in [k]}, M(S, X) \mid V, Z) = I((̂T^{(i)}, X^{(i)})_{i \in [k]}, M(S, X) \mid Z) = I((̂T^{(i)}, X^{(i)})_{i \in [k]}, M(S, X) \mid Z, T = ̂T) = \text{IIC}_{κ_4}(Π \mid T = ̂T),$$

where the second equality above follows from the fact that $(S, X)$ and $V$ are conditionally independent given $Z$, and the last equality follows from the fact that $(X, Z) \sim μ_ε$ in Protocol 4.

To sum up, the one-way protocol $Π'$ computes $\text{HD}_k$ with error probability at most $O(θ)$ and has information cost $o(k)$ bits (by Properties 3 and (b) above). By averaging over the shared randomness, we can convert $Π'$ into a one-way private-coin protocol $Π''$ with the same error and information cost guarantees. Note that since $U$ and $V$ are independent, we have that $I(U, V; M''(U)) = I(U; M''(U) \mid V)$ where $M''$ is the single message sent from Alice to Bob under $Π''$. Applying the generic compression result of [JRS03] (i.e., Lemma 5.7.8) with error parameter $ζ = θ$, we get that there exists a one-way deterministic protocol $Π''$ that computes $\text{HD}_k$ with error probability at most $O(θ)$ over the uniform distribution and with communication cost $o(k)$ bits. This contradicts the lower bound of Woodruff [Woo07] (i.e., Lemma 5.7.7).

$$\Box$$
Protocol 4 Simulation Protocol II’

Inputs. Alice is given $U \in \{0, 1\}^k$ and Bob is given $V \in \{0, 1\}^k$.

Parameters. A fixed sequence $T \triangleq (T^{(i)} \subseteq [n])_{i \in [k]}$ of subsets and noise parameters $\epsilon, q > 0$.

1. Alice and Bob use their shared randomness to sample $Z \in R \{0, 1\}^{k \cdot n}$.

2. Alice uses her private randomness to sample $X \in \{0, 1\}^{k \cdot n}$ to be an $\epsilon$-noisy copy of $Z$ conditioned on $(\langle T^{(i)}, X^{(i)} \rangle)_{i \in [k]} = U$.

3. Bob uses his private randomness to sample $Y \in \{0, 1\}^{k \cdot n}$ to be an $\epsilon$-noisy copy of $Z$ conditioned on $(\langle T^{(i)}, Y^{(i)} \rangle)_{i \in [k]} = V$.

4. Alice uses her private randomness to sample a sequence $S \triangleq (S^{(i)} \subseteq [n])_{i \in [k]}$ of subsets which is set to be a $q$-noisy copy of $T$.

5. Alice and Bob simulate the one-way deterministic protocol $\Pi$ on inputs $((S, X), (T, Y))$ and return the resulting output.

5.7.5 Proof of Lemma 5.7.3

Assume for the sake of contradiction that there is a deterministic one-way protocol $\Pi$ computing $F$ with error at most $\theta$ over the distribution $\nu_\epsilon$, and that has communication cost $o(k \cdot \epsilon \cdot n)$ bits. Let $M \triangleq M(X, S)$ be the single message that is sent from Alice to Bob under $\Pi$. By the Simulation Lemma 5.7.10, we should have that

$$I_{((S, X), (T, Y)) \sim \kappa} \left( (\langle T^{(i)}, X^{(i)} \rangle)_{i \in [k]}; M(X, S) \mid Y, T \right) \geq \beta \cdot k \quad (5.21)$$

for some constant $\beta$ that only depends on $\theta$.

By Lemma 5.7.9 and Equation (5.21), there exists a deterministic function $E(Y, T, M(X, S)) \in \{0, 1\}^k$ such that $E(Y, T, M(X, S)) = (\langle T^{(i)}, X^{(i)} \rangle)_{i \in [k]}$ with probability at least $(0.5 + \gamma)^k$ for some positive constant $\gamma$ that only depends on $\theta$. Hence, by applying the function $E(Y, T, M(X, S))$ to his inputs $(Y, T)$ and to the message $M(X, S)$ that he receives from Alice, Bob can guess the sequence $(\langle T^{(i)}, X^{(i)} \rangle)_{i \in [k]}$ with probability $(0.5 + \gamma)^k$. By Corollary 5.7.6 – which combines the strong direct product theorem for discrepancy of [LSS08] (i.e., Lemma 5.7.4) and the base lower bound of [GKKS16] that was proved using the discrepancy method (i.e., Lemma 5.7.5) – we conclude that the protocol $\Pi$ should have communication cost $\Omega(k \cdot \epsilon \cdot n)$ bits.
5.8 Lower Bound on Private-Coin Uncertain Protocols

In this section, we prove Theorem 5.1.11. We start by giving a high-level overview before giving the proof.

5.8.1 Overview of Proof of Theorem 5.1.11

As will be explained in detail below, the proof of Theorem 5.1.11 is based on a construction that requires us to understand the following “subset-majority with side information” setup. Alice is given a subset \( S \subseteq [n] \) and a string \( x \in \{\pm 1\}^n \), and Bob is given a subset \( T \subseteq [n] \) and a string \( y \in \{\pm 1\}^n \). The subsets \( S \) and \( T \) are adversarially chosen but are promised to satisfy \( S \subseteq T \), \( |T| = \ell \) and \( |T \setminus S| \leq \delta \cdot \ell \) for some fixed parameters \( \ell \) and \( \delta \). The strings \( x \) and \( y \) are chosen independently and uniformly at random. Alice and Bob wish to compute the function \( \text{SubsetMaj}((S, x), (T, y)) \triangleq \text{Sign}\left(\sum_{i \in T} x_i y_i\right) \). In words, \( \text{SubsetMaj}((S, x), (T, y)) \) is equal to 0 if \( x \) and \( y \) differ on a majority of the coordinates in subset \( T \), and 1 otherwise.

Note that \( S \) does not directly appear in the definition of the function \( \text{SubsetMaj} \) but it can serve as useful side-information for Alice.

What is the private-coin communication complexity of computing \( \text{SubsetMaj} \) on every \((S, T)\)-pair satisfying the above promise and with high probability over the random choice of \((x, y)\) and over the private randomness? We prove the following (informally stated) lower bound.

**Lemma 5.8.1.** Any private-coin protocol computing \( \text{SubsetMaj} \) on every \((S, T)\)-pair satisfying the promise and with high probability over the random choice of \((x, y)\) and over the private randomness should communicate at least \( \log(O(n)) \) bits for some positive integer \( t \) that depends on \( \delta \) and the error probability.

We next explain how the proof of Theorem 5.1.11 leads to the setup of Lemma 5.8.1 and how we prove Lemma 5.8.1.

**Reduction to Lemma 5.8.1.** In order to prove Theorem 5.1.11, we need to devise a function class for which circumventing the uncertainty is much easier using public randomness than using private randomness. One general setup in which Bob can leverage public randomness to resolve some uncertainty regarding Alice’s knowledge is the following “small-set intersection” problem. Assume that Alice is given a subset \( S \subseteq [n] \), and Bob is given a subset \( T \subseteq [n] \) such that \( T \) contains \( S \) and \( |T| = \ell \), where we think of \( \ell \) as being a large constant. Here, Bob knows that Alice has a subset of his own \( T \) but he is uncertain which subset Alice has. Using public randomness, a standard one-way hashing protocol communicating \( O(\ell) \) bits allows Bob to determine \( S \) with high probability. On the other hand, using only private randomness, the communication complexity of this task is \( \Theta(\log \log n) \) bits.

With the above general setup in mind, we consider functions \( f_S \) indexed by small subsets \( S \) of coordinates on which they depend. Since we want the functions \( f_S \) and \( f_T \) to be close
in Hamming distance, we enforce $|T \setminus S|$ to be small for every pair $(f_S, f_T)$ of functions in our class, and we let each function $f_S$ be “noise-stable”. Since we want our function $f_S$ to genuinely depend on all coordinates in $S$, the majority function $f_S(x, y) = \text{Sign}(\sum_{i \in S} x_i y_i)$ for $x, y \in \{\pm 1\}^n$ arises as a natural choice. We also let $x$ and $y$ be independent uniform-random strings. In this case, it can be seen that if $|T \setminus S|$ is a small constant fraction of $|T|$, then the quadratic polynomials $\sum_{i \in S} x_i y_i$ and $\sum_{i \in T} x_i y_i$ behave like standard Gaussians with correlation close to 1, and the quadratic threshold functions $f_S(x, y)$ and $f_T(x, y)$ are thus close in Hamming distance.

Note that in the certain case, i.e., when both Alice and Bob agree on $S$, they can easily compute $f_S(x, y)$ by having Alice send to Bob the $\ell$ bits $(x_i)_{i \in S}$. Moreover, if Alice and Bob are given access to public randomness in the uncertain case, Bob can figure out $S$ via the hashing protocol mentioned above using $\tilde{O}(\ell)$ bits of communication, which would reduce the problem to the certain case\footnote{Alternatively, Alice and Bob can run the protocol in Corollary 5.1.8 which would communicate $O(\ell)$ bits.}. The bulk of the proof will be to lower-bound the private-coin uncertain communication. Note that by the choice of our function class and distribution, this is equivalent to proving Lemma 5.8.1.

**Proof of Lemma 5.8.1.** To prove Lemma 5.8.1, the high-level intuition is that a protocol solving the uncertain problem should be essentially revealing to Bob the subset $S$ that Alice holds. Formalizing this intuition turns out to be challenging, especially that a private-coin protocol solving the uncertain problem is only required to output a single bit which is supposed to equal the Boolean function $f_T(x, y)$ with high probability over $(x, y)$ and over the private randomness. In fact, this high-level intuition can be shown not to hold in certain regimes\footnote{For example, for constant error probabilities, the one-way randomized communication complexity of small-set intersection is known to be $\Theta(\ell \cdot \log(\ell))$ bits (see, e.g., [BCK+14]) whereas the public-coin protocol in Corollary 5.1.8 can compute $f_T$ with $O(\ell)$ bits of communication.}. Moreover, the standard proofs that lower-bound the communication complexity of small-set intersection do not extend to lower-bound the communication complexity of $f_T$.

To lower-bound the private-coin communication of solving the uncertain task by a growing function of $n$, we consider the following shift communication game. Bob is given a sorted tuple $\sigma = (\sigma_1, \ldots, \sigma_t)$ of integers with $1 \leq \sigma_1 < \cdots < \sigma_t \leq n$, and Alice is either given the prefix $(\sigma_1, \ldots, \sigma_{t-1})$ of length $t-1$ of $\sigma$ or the suffix $(\sigma_2, \ldots, \sigma_t)$ of length $t-1$ of $\sigma$. Bob needs to determine the input of Alice. We show that a celebrated lower bound of Linial [Lin92] on the chromatic number of certain related graphs implies a lower bound of $\log^{t+1}(n)$ on the private-coin communication of the shift communication game. We then show that any private-coin protocol solving the uncertain task can be turned into a private-coin protocol solving the shift-communication game with a constant (i.e., independent of $n$) blow-up in the communication (see Protocol 5).
5.8.2 Proof of Theorem 5.1.11

We start by formally describing the construction that is used to prove Theorem 5.1.11. Each function in our universe is specified by a subset $S \subseteq [n]$ and is of the form $f_S : \{\pm 1\}^n \times \{\pm 1\}^n \to \{0, 1\}$ with $f_S(X, Y) \triangleq \operatorname{Sign}(\sum_{i \in S} X_i Y_i)$ for all $X, Y \in \{\pm 1\}^n$. The function class is then defined by

$$\mathcal{F}_\delta \triangleq \{(f_S, f_T) : S \subseteq T, |T| = \ell \text{ and } |T \setminus S| \leq \delta' \cdot \ell\},$$

where $\delta' = \alpha \cdot \delta^2$ for some sufficiently small positive absolute constant $\alpha$, and $\ell = \ell(\delta)$ is a sufficiently large function of $\delta$. The input pair $(X, Y)$ is drawn from the uniform distribution on $\{\pm 1\}^{2n}$. We start with the proof of Part (i) of Theorem 5.1.11. It essentially follows from the fact that the polynomials $\sum_{i \in S} X_i Y_i$ and $\sum_{i \in T} X_i Y_i$ behave like zero-mean Gaussians with unit-variance and correlation $\sqrt{1 - \delta'}$.

We will need the following well-known fact which follows from Sheppard’s formula [She99].

**Fact 5.8.2.** If $(X, Y)$ is a pair of zero-mean Gaussians with correlation $\mathbb{E}[XY] = \rho$, then

$$\operatorname{Pr}[\operatorname{Sign}(X) \neq \operatorname{Sign}(Y)] = \frac{\arccos(\rho)}{\pi}.$$

We now prove Part (i) of Theorem 5.1.11.

**Proof of Part (i) of Theorem 5.1.11.** Let $S \subseteq T \subseteq [n]$ be such that $|T| = \ell$ and $|T \setminus S| \leq \delta' \cdot \ell$. For fixed $\ell$, the distance $\delta_\ell(f_S, f_T)$ decreases when $|T \setminus S|$ decreases. So it suffices to upper-bound $\delta_\ell(f_S, f_T)$ when $|T \setminus S| = \delta' \cdot \ell$. Assume that the coordinates in $T$ are $1 \leq t_1 < t_2 < \cdots < t_{\ell}$. Then, we define the random vectors $X', Y' \in \{0, 1\}^\ell$ as $X_i' = X_{t_i} Y_i$ for all $i \in [\ell]$, and $Y_i' = X_{t_i} Y_i$ if $t_i \in S$ and $Y_i' = 0$ if $t_i \notin S$. We will apply the two-dimensional Berry-Esseen Theorem 5.9.1 to $(X', Y')$. To do so, first note that the random pairs $(X_1', Y_1'), (X_2', Y_2'), \ldots, (X_{\ell}', Y_{\ell}')$ are independent. Moreover, for every $i \in [\ell]$ such that $t_i \in S$, the covariance matrix of $(X_i, Y_i)$ is given by $\Sigma_i = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. On the other hand, for $i \in [\ell]$ such that $t_i \notin S$, the covariance matrix of $(X_i, Y_i)$ is given by $\Sigma_i = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. Hence, the average (across the $\ell$ coordinates) covariance matrix is given by

$$\Sigma = \ell^{-1} \cdot \sum_{i \in [\ell]} \Sigma_i = \begin{bmatrix} 1 - \delta' & 1 - \delta' \\ 1 - \delta' & 1 - \delta' \end{bmatrix}.$$

The smallest and largest eigenvalues of $\Sigma$ are respectively given by

$$\lambda \triangleq \frac{1 - \delta'}{2} - \frac{\sqrt{5 \cdot (1 - \delta')^2 - 2 \cdot (1 - \delta') + 1}}{2} + \frac{1}{2},$$

$$\Lambda \triangleq \frac{1 - \delta'}{2} + \frac{\sqrt{5 \cdot (1 - \delta')^2 - 2 \cdot (1 - \delta') + 1}}{2} + \frac{1}{2}.$$
In Equation (5.22a), it can be checked that for \( \delta' \in (0, 1) \), \( \lambda > 0 \). By the two-dimensional Berry-Esseen Theorem 5.9.1, we get that
\[
\delta_U(f_S, f_T) = \Pr[\text{Sign}(\sum_{i \in S} X_i Y_i) \neq \text{Sign}(\sum_{i \in T} X_i Y_i)]
\]
\[
= \Pr[\text{Sign}(\sum_{i \in [\ell]} X'_i) \neq \text{Sign}(\sum_{i \in [\ell]} Y'_i)]
\]
\[
= \Pr[\text{Sign}(X'') \neq \text{Sign}(Y'')] \pm O\left(\frac{1}{\lambda^{3/2} \sqrt{\ell}}\right),
\]  
(5.23)
where \((X'', Y'')\) is a pair of zero-mean Gaussians with covariance matrix \(\Sigma\). We can scale \(Y''\) to make it have unit-variance; this does not change its mean or the probability in Equation (5.23). The covariance matrix becomes: \(\Sigma' = \begin{bmatrix} 1 & \sqrt{1 - \delta'} \\ \sqrt{1 - \delta'} & 1 \end{bmatrix}\). By Sheppard’s formula (i.e., Fact 5.8.2), we deduce that
\[
\delta_U(f_S, f_T) = \frac{\arccos(\sqrt{1 - \delta'})}{\pi} \pm O\left(\frac{1}{\lambda^{3/2} \sqrt{\ell}}\right)
\]
\[
= O(\sqrt{\delta'})
\]
\[
\leq \delta,
\]
where we have used the facts that \(\arccos(1 - x) = O(\sqrt{x})\) for small positive values of \(x\), that \(\delta' = \alpha \cdot \delta^2\) for a sufficiently small positive absolute constant \(\alpha\), and that \(\ell\) is be a sufficiently large function of \(\delta\).

We now give the proof of Part (ii) of Theorem 5.1.11, which is quite immediate.

**Proof of Part (ii) of Theorem 5.1.11.** When both Alice and Bob know the subset \(T \subseteq [n]\) which satisfies \(|T| \leq \ell\), Alice can send the sequence \((X_j : j \in T)\) of at most \(\ell\) bits to Bob who can then output \(f_T(X, Y)\).

To prove Part (iii) of Theorem 5.1.11, the next definition – which is based on the graphs studied by Linial [Lin92]– will be crucial to us.

**Definition 5.8.3 (Shift Communication Game \(\mathcal{G}_{m,t}\)).** Let \(m\) and \(t\) be positive integers with \(t \leq m\). In the communication problem \(\mathcal{G}_{m,t}\), Bob is given a sorted tuple \(\sigma = (\sigma_1, \ldots, \sigma_t)\) of distinct integers with \(1 \leq \sigma_1 < \cdots < \sigma_t \leq m\). In the YES case, Alice is given the prefix \((\sigma_1, \ldots, \sigma_{t-1})\) of length \(t - 1\) of \(\sigma\). In the NO case, Alice is given the suffix \((\sigma_2, \ldots, \sigma_t)\) of length \(t - 1\) of \(\sigma\). Alice and Bob need to determine which of the YES and NO cases occurs.

Lemma 5.8.4 lower-bounds the private-coin communication complexity of \(\mathcal{G}_{m,t}\). Its proof uses Linial’s lower bound on the chromatic number of related graphs.

**Lemma 5.8.4.** There is an absolute constant \(c\) such that for every sufficiently small \(\epsilon > 0\), we have that \(\text{PrivCC}_\epsilon(\mathcal{G}_{m,t}) \geq c \cdot \log^{(t+2)}(m)\).
We prove Lemma 5.8.4 in Section 5.8.3. The proof of Part (iii) of Theorem 5.1.11 – which is the main part in the proof of Theorem 5.1.11 – is given in Section 5.8.4.

5.8.3 Proof of Lemma 5.8.4

The following family of graphs was first studied by Linial (in the setup of distributed graph algorithms) [Lin92].

Definition 5.8.5 (Shift Graph $G_{m,t}$). Let $m$ and $t$ be positive integers with $t \leq m$. In the graph $G_{m,t} = (V_{m,t}, E_{m,t})$, the vertices are all sorted tuples $\sigma = (\sigma_1, \ldots, \sigma_t)$ of distinct integers with $1 \leq \sigma_1 < \cdots < \sigma_t \leq m$. Two such tuples $\sigma$ and $\pi$ are connected by an edge in $E_{m,t}$ if and only if either $(\sigma_1, \ldots, \sigma_{t-1}) = (\pi_2, \ldots, \pi_t)$ or $(\sigma_2, \ldots, \sigma_t) = (\pi_1, \ldots, \pi_{t-1})$.

Recall that the chromatic number $\chi(G)$ of an undirected graph $G$ is the minimum number of colors needed to color its vertices such that no two adjacent vertices share the same color. The following theorem is due to Linial.

Theorem 5.8.6 ([Lin92], Proof of Theorem 2.1). Let $m$ and $t$ be positive integers such that $t \leq m$ and $t$ is odd. Then, $\chi(G_{m,t}) \geq \log^{(t-1)}(m)$.

The next lemma uses Theorem 5.8.6 to lower-bound the deterministic two-way communication complexity of the shift communication game $G_{m,t}$.

Lemma 5.8.7. Let $m$ and $t$ be positive integers such that $t \leq m$ and $t$ is odd. Then, it is the case that $\text{CC}(G_{m,t}) \geq \log^{(t+1)}(m)$.

Proof of Lemma 5.8.7. Assume for the sake of contradiction that there exists a deterministic two-way protocol that computes $G_{m,t}$ and that has communication cost smaller than $\log^{(t+1)}(m)$. Then, using the fact that the one-way communication complexity of any function is at most exponential in its two-way communication complexity, we get that there is a one-way protocol $\Pi$ that computes $G_{m,t}$ and that has communication cost smaller than $\log^{(t)}(m)$. Let $M$ be the single message sent from Alice to Bob under $\Pi$. Then, the length of $M$ satisfies $|M| < \log^{(t)}(m)$. Note that Alice’s input is an element of the vertex-set $V_{m,t}$ of the shift graph $G_{m,t}$ (Definition 5.8.5). Since $M$ is a deterministic function of Alice’s input, it induces a coloring of $V_{m,t}$ into less than $2^{\log^{(t)}(m)} = \log^{(t-1)}(m)$ colors. The fact that no two adjacent vertices in $G_{m,t}$ share the same color follows from the correctness of $\Pi$ in computing $G_{m,t}$. This contradicts Theorem 5.8.6.

The following known fact gives a generic lower bound on the bounded-error private-coin communication complexity in terms of the deterministic communication complexity.

Fact 5.8.8 ([KN97], Theorem 3.14). For every communication function $f$ and every non-negative $\epsilon$ that is bounded below 1/2, we have that $\text{PrivCC}_\epsilon(f) = \Omega(\log(\text{CC}(f)))$.

Lemma 5.8.4 now follows by combining Lemma 5.8.7 and Fact 5.8.8.
5.8.4 Proof of Part (iii) of Theorem 5.1.11

In this section, we prove Part (iii) of Theorem 5.1.11. We assume for the sake of contradiction that there exists a one-way private-coin protocol $\Pi$ computing $\mathcal{F}_\delta$ with respect to the uniform distribution on $\{0,1\}^{2^n}$ with error at most $\epsilon/2 - 2\delta - \eta$ and with communication cost $|\Pi| = o(\eta^2 \cdot \log(t)(n))$ for some positive integer $t = \Theta((\epsilon/\delta)^2)$ that will be exactly specified later on. We will use $\Pi$ to give a one-way private-coin protocol $\Pi'$ solving the shift communication game $\mathcal{G}_{m,t}$ with high constant probability and with communication cost $|\Pi'| \leq O(\eta^2 \cdot |\Pi|) = o(\log(t)(n))$, which would contradict Lemma 5.8.4.

Description of Protocol $\Pi'$

The operation of protocol $\Pi'$, which uses $\Pi$ as a black-box, is described in Protocol 5. Note that the parameters in Protocol 5 are defined in terms of $\epsilon$ (which is a number in $[\delta, 0.5]$ that is given in the statement of Theorem 5.1.11) and $\delta'$ (which, as mentioned above, is set to $\alpha \cdot \delta^2$ for a sufficiently small positive absolute constant $\alpha$). Also as above, $\ell$ is set to a sufficiently large function of $\delta$. Note that $\epsilon' = O(\epsilon^2)$, and hence $t = O((\epsilon/\delta)^2)$. In Protocol 5, Bob is given as input a sorted tuple $\sigma$, and Alice is given as input either the prefix $\phi$ of $\sigma$ or its suffix $\psi$. In steps 1 and 2, Alice and Bob “stretch” their tuples, which amounts to each of them repeating each bit of the corresponding 0/1 indicator vector a certain number of times and appending a certain number of zeros (see Definition 5.8.9 and Figure 5-3 in Section 5.8.4 below for a thorough definition of stretch $r_m$). The aim of steps 1 and 2 is for Alice and Bob to produce a pair of subsets $(S, T)$ of $[n]$ such that $(f_S, f_T) \in \mathcal{F}_\delta$. The goal of Protocol 5 is for Bob to determine if Alice was given the prefix or the suffix of his tuple $\sigma$. To do so, Alice and Bob sample (using their private coins) $k$ random inputs $(X^{(i)}, Y^{(i)})_{i \in [k]}$ (steps 3 and 4) and then simulate the given private-coin protocol $\Pi$ to compute the function $f_T(X, Y)$ on the $k$ random input-pairs that were sampled (steps 6 and 7). Moreover, Alice sends to Bob the influential bits of $X^{(i)}$ for each $i \in [k]$ (step 8). The main idea will be to be for Bob to compute the empirical error corresponding to each of the prefix (step 11) and suffix (step 12), and then output the hypothesis with the smallest empirical error (step 13).
Protocol 5 Reduction Protocol Π′

Parameters. \( \epsilon' = 1 - \cos(\epsilon \pi), t = \lceil \epsilon'/\delta' \rceil, \quad r = \delta' \cdot \ell, \quad a = \ell \cdot (1 - t \cdot \delta'), \quad s = \ell \cdot (1 - \delta'), \quad k = \Theta(1/\eta^2) \).

Inputs. Bob is given a sorted tuple \( \sigma = (\sigma_1, \ldots, \sigma_t) \) of integers with \( 1 \leq \sigma_1 < \cdots < \sigma_t \leq (n - a)/r \). Alice is given a sorted tuple \( \lambda \in \{\phi, \psi\} \) where \( \phi \triangleq (\sigma_1, \ldots, \sigma_{t-1}) \) and \( \psi \triangleq (\sigma_2, \ldots, \sigma_t) \).

1. Alice sets \((\Lambda, S) \leftarrow \text{stretch}_{r, a}(\lambda)\).
2. Bob sets \((\Sigma, T) \leftarrow \text{stretch}_{r, a}(\sigma)\).
3. Alice uses her private randomness to sample \( k \) i.i.d. strings \( X^{(1)}, \ldots, X^{(k)} \in_R \{0, 1\}^n \).
4. Bob uses his private randomness to sample \( k \) i.i.d. strings \( Y^{(1)}, \ldots, Y^{(k)} \in_R \{0, 1\}^n \).
5. For \( i = 1, \ldots, k \):
   6. Alice and Bob simulate the protocol \( \Pi \) on inputs \(((S, X^{(i)}), (T, Y^{(i)}))\).
   7. Bob computes the resulting output bit \( B_i \).
   8. Alice sends to Bob the sequence of bits \((X_{\Lambda_1}^{(i)}, X_{\Lambda_2}^{(i)}, \ldots, X_{\Lambda_s}^{(i)})\).
9. EndFor
10. Bob sets \( \Phi \leftarrow \text{stretch}_{r, a}(\phi) \) and \( \Psi \leftarrow \text{stretch}_{r, a}(\psi) \).
11. Bob computes the “prefix error” \( \text{err}^{(p)} \triangleq \sum_{i \in [k]} \mathbb{1} \left[ \text{Sign} \left( \sum_{j \in [s]} X_{\Lambda_j}^{(i)} Y_{\Phi_j}^{(i)} \right) \neq B_i \right] \).
12. Bob computes the “suffix error” \( \text{err}^{(s)} \triangleq \sum_{i \in [k]} \mathbb{1} \left[ \text{Sign} \left( \sum_{j \in [s]} X_{\Lambda_j}^{(i)} Y_{\Psi_j}^{(i)} \right) \neq B_i \right] \).
13. Bob returns YES if \( \text{err}^{(p)} \leq \text{err}^{(s)} \) and NO otherwise.

The \text{stretch}_{r,a} Procedure

In this section, we thoroughly define and illustrate the stretching procedure used in Protocol 5 and mentioned in Section 5.8.

Definition 5.8.9 (The \text{stretch}_{r,a} Procedure). Let \( d, t, r \) and \( a \) be positive integers with \( t \leq d \). For any sorted tuple \( \sigma = (\sigma_1, \ldots, \sigma_t) \) of distinct integers with \( 1 \leq \sigma_1 < \cdots < \sigma_t \leq d \), we first let \( z \in \{0, 1\}^d \) be the 0/1 indicator vector of the subset of \([d]\) corresponding to \( \sigma \). Let \( w \in \{0, 1\}^{d \cdot r + a} \) be the string obtained from \( z \) by repeating each of its coordinates \( r \) times, and then appending \( a \) ones. Namely, for each \( i \in [d] \) and \( j \in [r] \), we set \( w_{(i-1) \cdot r + j} = z_i \) and
for each $j \in [a]$, we set $w_{d-r+j} = 1$. Then, we let $1 \leq \Sigma_1 < \Sigma_2 < \cdots < \Sigma_{t_r+a} \leq d \cdot r + a$ be the indices of the coordinates of $w$ that are equal to 1. The output of $\text{Stretch}_{r,a}(\sigma)$ is then the pair $(\Sigma, W)$ where $\Sigma \triangleq (\Sigma_1, \Sigma_2, \ldots, \Sigma_{t_r+a})$ and $W \subseteq [d \cdot r + a]$ is the support of $w$.

The operation of $\text{Stretch}_{r,a}$ is illustrated in Figure 5-3 in the particular case where $r = 2$, $a = 3$ and $d = 9$. The “appending parameter” $a$ allows us to control how small is the normalized Hamming distance between the binary strings corresponding to $\Sigma$ and $\Phi$ (respectively $\Psi$). The purpose of the “repetition parameter” $r$ is the following. Consider the tuples $\phi = (2, 4, 5, 7)$ and $\psi = (4, 5, 7, 9)$ in Figure 5-3. The number of differing coordinates between the tuples $\phi$ and $\psi$ is 4. After stretching, the number of differing coordinates between the resulting tuples $\Phi$ and $\Psi$ is amplified to $4 \cdot r = 8$. These two notions of distance (the number of coordinates on which the tuples differ and the normalized Hamming distance between the corresponding binary strings) are important to us. The fact that these two distances are important to us is the reason why we let the stretch algorithm have two equivalent outputs (a subset $W$ and a tuple $\Sigma$).

Bob’s subset: $0\ 1\ 0\ 1\ 1\ 0\ 1\ 0\ 1$ \hspace{1cm} $\text{stretch}_{r,a}$ \hspace{1cm} $0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 1$  

$\sigma = (2, 4, 5, 7, 9)$ \hspace{2cm} $\Sigma = (3, 4, 7, 8, 9, 10, 13, 14, 17, 18, 19, 20, 21)$

Prefix of $\sigma$: $0\ 1\ 0\ 1\ 1\ 0\ 1\ 0$ \hspace{1cm} $\text{stretch}_{r,a}$ \hspace{1cm} $0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1$  

$\phi = (2, 4, 5, 7)$ \hspace{2cm} $\Phi = (3, 4, 7, 8, 9, 10, 13, 14, 19, 20, 21)$

Suffix of $\sigma$: $0\ 0\ 0\ 1\ 1\ 0\ 1\ 0$ \hspace{1cm} $\text{stretch}_{r,a}$ \hspace{1cm} $0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1$  

$\psi = (4, 5, 7, 9)$ \hspace{2cm} $\Psi = (7, 8, 9, 10, 13, 14, 17, 18, 19, 20, 21)$

Figure 5-3: Operation of Stretch$_{r,a}$ Procedure for $r = 2$ and $a = 3$.

Analysis of Protocol $\Pi'$

We now turn to the formal analysis of Protocol 5. First, note that the communication cost of protocol $\Pi'$ satisfies $|\Pi'| \leq O(\eta^{-2} \cdot |\Pi|)$. Let $\Lambda$, $\Sigma$, $\Phi$ and $\Psi$ be the ordered sequences defined in the operation of Protocol 5. Define the functions $g$ and $h$ as

$$g(X, Y) \triangleq \text{Sign}\left( \sum_{j \in [s]} X_{\Lambda_j} Y_{\Phi_j} \right), \quad (5.24a)$$

$$h(X, Y) \triangleq \text{Sign}\left( \sum_{j \in [s]} X_{\Lambda_j} Y_{\Psi_j} \right). \quad (5.24b)$$

Note that steps 11 and 12 of the protocol compute the empirical errors of functions $g$ and $h$ respectively. Since $\Lambda \in \{\Phi, \Psi\}$, let’s assume without loss of generality that $\Lambda = \Phi$ and show that the protocol $\Pi'$ returns YES with high probability. The case where $\Lambda = \Psi$ is symmetric. When $\Lambda = \Phi$, we have that $g = f_S$ where $S$ is the subset of $[n]$ that Alice gets in step 1. The operation of the stretch algorithm (Definition 5.8.9) guarantees that $S \subseteq [n]$, $|T| = \ell$ and $|T \setminus S| = \delta' \cdot \ell$. Part (i) of Theorem 5.1.11 then implies that $\delta_{\ell}(g, f_T) \leq \delta$. In
order to show that in step 13 Bob returns YES with high probability, the main idea will be to lower bound the distance between the functions $g$ and $h$. This is done in the next lemma.

**Lemma 5.8.10.** The functions $g$ and $h$ defined in Equations (5.24a) and (5.24b) satisfy$$\delta_{\mathcal{U}}(g, h) \geq \epsilon - \delta.$$

To prove Lemma 5.8.10, we use the next lemma which spells out the distribution of the sequence $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})_{j \in [s]}$ of random variables.

**Lemma 5.8.11.** The random variables $(X_{\Lambda_1}Y_{\Phi_1}, X_{\Lambda_1}Y_{\Psi_1}), (X_{\Lambda_2}Y_{\Phi_2}, X_{\Lambda_2}Y_{\Psi_2}), \ldots, (X_{\Lambda_s}Y_{\Phi_s}, X_{\Lambda_s}Y_{\Psi_s})$ are independent, and they are distributed as follows:

1. For $1 \leq j \leq s - a$, $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is uniformly distributed on $\{\pm 1\}^2$.
2. For $s - a < j \leq s$, $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is uniformly distributed on $\{(-1, -1), (+1, +1)\}$.

**Proof of Lemma 5.8.11.** By the operation of the Stretch$_{r,a}$ procedure (Definition 5.8.9) in steps 1 and 2 of Protocol 5, for every $s - a < j \leq s$, it holds that $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j}) = (X_jY_j, X_jY_j)$ and that $X_j$ and $Y_j$ do not contribute to any other $(X_{\Lambda_j}Y_{\Phi_j}', X_{\Lambda_j}Y_{\Psi_j}')$ pair. This implies part 1 and that for each $s - a < j \leq s$, the pair $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is independent of all other pairs in the sequence.

We next prove by induction on $1 \leq j \leq s - a$ that conditioned on $(X_{\Lambda_{<j}}Y_{\Phi_{<j}}, X_{\Lambda_{<j}}Y_{\Psi_{<j}})$ taking any particular value, $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is uniformly distributed on $\{\pm 1\}^2$. To see this (assuming without loss of generality that $\Lambda = \Phi$), note that $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j}) = (X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ where $X_{\Lambda_j}$ and $Y_{\Psi_j}$ do not appear in $(X_{\Lambda_{<j}}Y_{\Phi_{<j}}, X_{\Lambda_{<j}}Y_{\Psi_{<j}})$. Hence, conditioned on $(X_{\Lambda_{<j}}Y_{\Phi_{<j}}, X_{\Lambda_{<j}}Y_{\Psi_{<j}})$ taking any particular value, $X_{\Lambda_j}Y_{\Phi_j}$ is a uniformly random bit (because of $X_{\Lambda_j}$). Moreover, conditioned on $(X_{\Lambda_{<j}}Y_{\Phi_{<j}}, X_{\Lambda_{<j}}Y_{\Psi_{<j}})$ and $X_{\Lambda_j}$, $Y_{\Phi_j}$ taking any particular values, $X_{\Lambda_j}Y_{\Psi_j}$ is a uniformly random bit (because of $Y_{\Psi_j}$). This completes the proof of the lemma.

We now use Lemma 5.8.11 along with a two-dimensional Central Limit Theorem in order to prove Lemma 5.8.10.

**Proof of Lemma 5.8.10.** By Lemma 5.8.11, independently for each $1 \leq j \leq s - a$, $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is a $\rho_j$-correlated random pair with $\rho_j = 0$, and independently for each $s - a < j \leq s$, $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ is a $\rho_j$-correlated random pair with $\rho_j = 1$. Hence, the average correlation across coordinates is

$$\rho \triangleq k^{-1} \cdot \sum_{j \in [k]} \rho_j = \frac{a}{s} = \frac{\ell \cdot (1 - t \cdot \delta')}{\ell \cdot (1 - \delta')} = 1 - \frac{(t - 1) \cdot \delta'}{1 - \delta'} \leq 1 - (\epsilon' - \delta'),$$

where the last inequality used the setting of $t = [\epsilon' / \delta']$ in Protocol 5. Denoting by $\Sigma_j$ the covariance matrix of $(X_{\Lambda_j}Y_{\Phi_j}, X_{\Lambda_j}Y_{\Psi_j})$ for every $j \in [s]$, the average (across coordinates) covariance matrix is then given by $\Sigma \triangleq k^{-1} \cdot \sum_{j \in [k]} \Sigma_j = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Note that the smallest
eigenvalue of $\Sigma$ is $\lambda = 1 - |\rho|$. Let $(X', Y')$ be a pair of zero-mean $\rho$-correlated Gaussians. By the two-dimensional Berry-Esseen Theorem 5.9.1 and Sheppard’s formula (i.e., Fact 5.8.2), we get that

$$
\delta_U(g, h) = \Pr[g(X, Y) \neq h(X, Y)]
= \Pr[g(X, Y) = 0, h(X, Y) = 1] + \Pr[g(X, Y) = 1, h(X, Y) = 0]
= \Pr[X' < 0, Y \geq 0] + O\left(\frac{1}{(1 - |\rho|)^{3/2} \cdot \sqrt{s}}\right)
= \Pr[\text{Sign}(X') \neq \text{Sign}(Y')] + O\left(\frac{1}{(1 - |\rho|)^{3/2} \cdot \sqrt{\ell}}\right)
= \frac{\arccos(\rho)}{\pi} + O\left(\frac{1}{(1 - |\rho|)^{3/2} \cdot \sqrt{\ell}}\right)
\geq \frac{\arccos(1 - (\epsilon' - \delta'))}{\pi} - O\left(\frac{1}{(1 - |\rho|)^{3/2} \cdot \sqrt{\ell}}\right)
\geq \frac{\arccos(1 - \epsilon')}{\pi} - O(\sqrt{\delta'}) - O\left(\frac{1}{(1 - |\rho|)^{3/2} \cdot \sqrt{\ell}}\right)
\geq \epsilon - \delta,
$$

where the last inequality follows by setting $\epsilon' = 1 - \cos(\epsilon \pi)$ for the given $\epsilon \in [\delta, 0.5]$, setting $\ell$ to be a sufficiently large function of $\delta$, and setting $\delta' = \alpha \cdot \delta^2$ for a sufficiently small positive absolute constant $\alpha$. \hfill \Box

We are now ready to complete the proof of Part (iii) of Theorem 5.1.11. In Protocol 5, the tuples $((X^{(i)}, Y^{(i)}), B_i)_{i \in [k]}$ are i.i.d. samples corresponding to a function $q$ which, by the error guarantee of protocol $\Pi$, is $(\epsilon/2 - 2\delta - \eta)$-close to $f_T$. Since $\delta_U(g, f_T) \leq \delta$, we get that $\delta_T(g, q) \leq (\epsilon/2 - \delta - \eta)$. By Lemma 5.8.10, we also get that $\delta_U(h, q) \geq (\epsilon/2 + \eta)$. By Hoeffding’s bound (i.e., Fact 5.8.12), for $k = \Theta(1/\eta^2)$, the empirical error $\epsilon_T^{(g)}$ of $g$ on the samples $((X^{(i)}, Y^{(i)}), B_i)_{i \in [k]}$ is less than the empirical error $\epsilon_T^{(h)}$ of $h$ on these samples, with high constant probability. Hence, Bob returns YES with high constant probability. A symmetric argument shows that when $\Lambda = \Psi$, Bob returns NO with high constant probability, which completes the proof.

**Fact 5.8.12** (Hoeffding’s bound). Consider a coin that shows up head with probability $p$. Let $H(k)$ be the number of heads obtained in $k$ independent tosses of this coin. Then, for every $\epsilon > 0$, $\Pr[|H(k) - pk| > \epsilon k] \leq 2e^{-2\epsilon^2 k}$.

**Remark 5.8.13.** As mentioned in Remark 5.1.13, the above construction cannot give a separation larger than $\Theta(\log \log n)$. This is because using private randomness, Bob can learn the set $S$ using $O(\log \log n)$ bits of communication (see, e.g., [BCK+14]). Additionally, Alice can send the coordinates of $X$ indexed by the elements of $S$ to Bob who can then compute $f_S(X, Y)$.
5.9 Two-Dimensional Berry-Esseen Theorem for Independent Random Variables

In this section, we state a two-dimensional Berry-Esseen theorem for independent (but not necessarily identically distributed) binary random variables that we used in the proofs in Section 5.8 (namely, in the proofs of Part (i) of Theorem 5.1.11 and of Lemma 5.8.10). It follows from a known multi-dimensional Berry-Esseen theorem and the argument is very similar to that of [MORS10], the only exceptions being that in our case the random variables are not necessarily identically distributed, and each of them takes values in \{-1, 0, +1\} instead of \{±1\}.

**Theorem 5.9.1 (Two-dimensional Berry-Esseen).** Consider the linear form \(\ell(z) \triangleq k^{-1/2} \sum_{i \in [k]} z_i\) where \(z \in \{-1, 0, +1\}^k\). Let \((x, y) \in \{-1, 0, +1\}^k \times \{-1, 0, +1\}^k\) such that independently for each \(i \in [k]\), \((x_i, y_i)\) is a pair of zero-mean random variables with covariance matrix \(\Sigma_i\). Let \(\Sigma \triangleq k^{-1} \sum_{i \in [k]} \Sigma_i\) and denote by \(\lambda\) the smallest eigenvalue of \(\Sigma\). Then, for any intervals \(I_1, I_2 \subseteq \mathbb{R}\), it holds that

\[
|\Pr[(\ell(x), \ell(y)) \in I_1 \times I_2] - \Pr[(X, Y) \in I_1 \times I_2]| \leq O\left(\frac{1}{\lambda^{3/2} \sqrt{k}}\right),
\]

where \((X, Y)\) is a pair of zero-mean Gaussians with covariance matrix \(\Sigma\).

**Proof.** The following statement appears as Theorem 16 in [KKMO07] and as Corollary 16.3 in [BR86].

**Theorem 5.9.2.** Let \(X_1, \ldots, X_k\) be independent random variables taking values in \(\mathbb{R}^d\) and satisfying:

- \(\mathbb{E}[X_j]\) is the all-zero vector for every \(j \in \{1, \ldots, k\}\).
- \(w^{-1} \sum_{j=1}^w \text{Cov}[X_j] = \Sigma\) where \(\text{Cov}\) denotes the covariance matrix.
- \(\lambda\) is the smallest eigenvalue of \(\Sigma\) and \(\Lambda\) is the largest eigenvalue of \(\Sigma\).
- \(\rho_3 = k^{-1} \sum_{j=1}^k \mathbb{E}[||X_j||^3] < \infty\).

Let \(Q_k\) denote the distribution of \(k^{-1/2}(X_1 + \cdots + X_k)\), let \(\Phi_{0,V}\) denote the distribution of the \(d\)-dimensional Gaussian with mean 0 and covariance matrix \(\Sigma\), and let \(\eta = C \lambda^{-3/2} \rho_3 k^{-1/2}\), where \(C\) is a certain universal constant. Then, for any Borel set \(A\),

\[
|Q_k(A) - \Phi_{0,V}(A)| \leq \eta + B(A),
\]

where \(B(A)\) is the following measure of the boundary of \(A\): \(B(A) = 2 \sup_{y \in \mathbb{R}^d} \Phi_{0,V}((\partial A)\eta' + y)\), \(\eta' = \Lambda^{1/2} \eta\) and \((\partial A)\eta'\) denotes the set of points within distance \(\eta'\) of the topological boundary of \(A\).
We now apply Theorem 5.9.2 with $d = 2$ in order to complete the proof of Theorem 5.9.1. We are given that for every $i \in \{1, \ldots, k\}$, $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = 0$ and $\text{Cov}(X_i, Y_i) = \Sigma_i$. Thus, $k^{-1} \sum_{j=1}^{k} \text{Cov}(X_j, Y_j) = \sum_{j=1}^{k} \Sigma_j = \Sigma$. Note that the largest eigenvalue of $\Sigma$ is $\Lambda = O(1)$. Moreover, since each coordinate of our random variables is $\{-1, 0, +1\}$-valued, for every $j \in \{1, \ldots, k\}$, $\mathbb{E}[||X_j||^3] \leq 2^{3/2}$. Thus, $\rho_3 \leq 2^{3/2}$. Hence, $\eta = O(\Lambda^{1/2} \eta)$. Since $\eta' = \Lambda^{1/2} \eta = O(\eta)$, Theorem 5.9.1 follows.

5.10 The Need to Work with Positive-Error Uncertain Protocols

For completeness, we exhibit a class $\mathcal{F}$ of pairs of Boolean-valued functions such that for every $(f, g) \in \mathcal{F}$, the functions $f$ and $g$ are very close with respect to the uniform distribution, the zero-error communication complexity of each of $f$ and $g$ in the standard model is a single bit, but the zero-error communication in the uncertain model is quite large. Formally, we prove the following:

**Theorem 5.10.1.** For every $\delta \in (0, 1)$, there exists a class $\mathcal{F}$ of pairs of Boolean-valued functions over the domain $\{0, 1\}^n$ such that

1. $\delta_\nu(\mathcal{F}) \leq \delta$.
2. $\text{CC}_0(\mathcal{F}) = 1$.
3. $\text{PubCCU}_0(\mathcal{F}) = \Omega(n - \log(1/\delta))$,

where $\nu$ is the uniform distribution on $\{0, 1\}^n \times \{0, 1\}^n$.

To prove Theorem 5.10.1, we will need the following lower bound on the well-studied INDEXING function. Recall that in the INDEXING$_m$ problem with parameter $m$, Alice is given an element $x \in [m]$, Bob is given a function $h: [m] \rightarrow \{0, 1\}$, and they are required to output $h(x)$.

The next theorem asserts that the two-way communication complexity of INDEXING$_m$ is $\Omega(\log m)$ bits. Note that this bound is essentially tight as Alice can send her input to Bob using $\log m$ bits of communication.

**Theorem 5.10.2.** There is a constant $\epsilon > 0$ such that

$$\text{CC}_\epsilon(\text{INDEXING}_m) = \Omega(\log m).$$

Theorem 5.10.2 follows from the well-known fact that the one-way communication complexity from Bob to Alice of INDEXING$_n$ (a.k.a., the “hard direction”) is $\Omega(n)$ bits [KN97,

---

The standard definition of INDEXING terms $x$ the “index” and views $h$ as a vector in $\{0, 1\}^m$. Our version is equivalent and a little more convenient notationally.

---

159
Exercise 4.20] and the generic fact that there is at most an exponential gap between one-way communication complexity and two-way communication complexity [KN97, Exercise 4.21].

We are now ready to prove Theorem 5.10.1.

**Proof of Theorem 5.10.1.** We first describe the class of functions we work with. Let $T \subseteq \{0,1\}^n$ be a set of size $\delta \cdot 2^n$. Let $\mathcal{F}' = \{g' : \{0,1\}^n \to \{0,1\} \mid g'(x) = 0, \forall x \not\in T\}$. We now define $\mathcal{F}$ in terms of $\mathcal{F}'$ as follows:

$$\mathcal{F} = \{(0,g) \mid \exists g' \in \mathcal{F}' \text{ s.t. } g(x,y) = g'(x), \forall (x,y)\}.$$ 

So the first function $f$ is always the 0 function, and the second function $g$ depends only on $x$ and is always 0 if $x \not\in T$.

Since $f \neq g$ only when $x \in T$ and this happens with probability $\delta$, we have

$$\delta_v(f,g) \overset{\Delta}{=} \Pr_{(X,Y) \sim \nu} [f(X,Y) \neq g(X,Y)] \leq \Pr_{(X,Y) \sim \nu} [X \in T] = \delta.$$

We conclude that $\delta_v(\mathcal{F}) \overset{\Delta}{=} \max_{(f,g) \in \mathcal{F}} \{\delta_v(f,g)\} \leq \delta$ yielding Part (1) of the Theorem. Part (2) is immediate from the fact that $g(x,y) = g'(x)$ for every $(f,g) \in \mathcal{F}$ and so, in the certain-communication setting, Alice can compute $g'(x)$ and send it to Bob.

We now turn to Part (3) for which we give a reduction from INDEXING$_m$ for $m = \delta \cdot 2^n$. Suppose PubCCU$(\mathcal{F}) \leq k$ and so there is a protocol $\Pi$ that communicates $k$ bits such that if Alice is given $(f,x)$ and Bob $(g,y)$ with $(f,g) \in \mathcal{F}$, the protocol outputs $g(x,y)$. We show $k = \Omega(\log m) = \Omega(n - \log(1/\delta))$. (Note that since $\Pi$ is a zero-error protocol, we have that the output of $\Pi$ is correct on all valid inputs, and so we can ignore the distribution on $(x,y)$ below.) Associate $[m]$ with the set $T$, so that Alice’s input is an element $x \in T$ and Bob’s input is a function $h : T \to \{0,1\}$, and their goal is to compute $h(x)$. Let $g' : \{0,1\}^n \to \{0,1\}$ be given by $g'(x) = 0$ if $x \not\in T$ and $g'(x) = h(x)$ otherwise. Let $g(x,y) = g'(x)$. Alice can map her input to the pair $(0,x)$ and Bob to the pair $(g,0)$, and now they have inputs to our uncertain communication problem on $\mathcal{F}$. Running $\Pi$ on this pair produces as output $g(x) = h(x)$ (since $x \in T$) which is the desired output of INDEXING$_m$. Applying 5.10.2 we conclude $k = \Omega(\log m)$ and this yields Part (3).

\[\square\]

### 5.11 Discussion and Future Directions

Functional uncertainty models much of the day-to-day interactions among humans, where a person is somewhat aware of the objectives of the other person she is interacting with, but do not know them precisely. Neither person typically knows exactly what aspects of their own knowledge may be relevant to the interaction, yet they do manage to have a short conversation. This is certainly a striking phenomenon that has been mostly unexplained in mathematical terms. This chapter initiated the exploration of such phenomena. It is important to understand what mechanisms may come into play here, and what features play a role. Is the ability to make random choices important? Is shared information crucial? Is
there a particular measure of distance between functions that makes efficient communication feasible? In order to understand such questions, one first needs to have a ground-level understanding of communication with functional uncertainty. This chapter tackled several basic questions in this setting, and it raises numerous interesting questions, some of which we summarize next.

On the technical side, it would be very interesting to determine the correct exponent of $I(X; Y)$ in Theorem 5.1.6. Theorem 5.1.6 and Theorem 5.1.3 imply that this exponent is between $1/2$ and 1. Moreover, it would be nice to understand the needed dependence on $k$ in the product $k \cdot I(X; Y)$ in Theorem 5.1.6. A related (but perhaps more challenging) question is whether the dependence on $n$ can be improved from $\Omega(\sqrt{n})$ to $\Omega(n)$ in Theorem 5.1.10 (while keeping the communication in the standard case equal to $O(1)$). As discussed in Section 5.3, such an improvement would require a new construction of a family of pairs of Boolean functions and an input distribution since the $\Omega(\sqrt{n})$ lower bound is tight (up to a logarithmic factor) for the considered construction.

An ideal protocol for communication amid uncertainty would only use private randomness (or even no randomness at all). The questions of determining the tight bounds for communication amid uncertainty in the deterministic, private-coin and imperfectly shared randomness setups remain open, and are likely to require fundamentally new ideas and constructions. For instance, can one prove a non-trivial upper bound – such as Theorem 5.1.6 – on the communication complexity of deterministic protocols?

As mentioned in Remark 5.1.13 and Remark 5.1.5 in Section 5.1, significantly improving the bounds in Theorems 5.1.11 and Theorem 5.1.4 seems to require fundamentally new constructions, and is a very important question. Moreover, is there an analogue of the protocol in Theorem 5.1.9 for non-product distributions?

It would also be extremely interesting to prove an analogue of Theorem 5.1.6 for two-way protocols. Our proof of Theorem 5.1.6 uses in particular the fact that any low-communication one-way protocol in the standard distributional communication model should have a canonical form: to compute $g(x, y)$, Alice tries to describe the entire function $g(x, \cdot)$ to Bob, and this does not create a huge overhead in communication. Coming up with a canonical form of two-way protocols that somehow changes gradually as we morph from $g$ to $f$ seems to be the essence of the challenge in extending Theorem 5.1.6 to the two-way setting. A concrete question here is whether the dependence on $k$ in the special case of product distributions (Equation (5.2) of Corollary 5.1.7 with $I(X; Y) = 0$) can be improved from $2^k$ to $\text{poly}(k)$. More simply, if $k$ is the $r$-round certain communication, can we upper bound the $r$-round uncertain communication by some function of $k$, $I$ and possibly $r$? Even for $r = 2$ and when the uncertain protocol is allowed to use public randomness, no protocols (other than the ones given in Corollary 5.1.7) are known in this setting. On the other hand, no separations are known for this case (beyond those known for $r = 1$) even if the protocols are restricted to be deterministic.

On the more conceptual side, arguably, the model considered here is realistic: communication has some goals which we model by letting Bob be interested in a specific function of the joint information that Alice and Bob possess. Moreover, it is an arguably natural
model to posit that the two are not in perfect synchronization about the function that Bob is interested in, but Alice can estimate the function in some sense. One aspect of our model that can be further refined is the specific notion of distance that quantifies the gap between Bob’s function and Alice’s estimate. We here chose the Hamming distance which forms a good first starting point. We believe that it is interesting to propose and study other models of distance between functions that more accurately capture natural forms of uncertainty.

Finally, we wish to emphasize the mix of adversarial and probabilistic elements in our uncertainty model — the adversary picks \((f, g)\) whereas the inputs \((X, Y)\) are sampled from a distribution. We believe that richer mixtures of adversarial and probabilistic elements could lead to broader settings of modelling and coping with uncertainty — the probabilistic elements offer efficient possibilities that are often immediately ruled out by adversarial choices, whereas the adversarial elements prevent the probabilistic assumptions from being too precise.
Chapter 6

Uncertain Distributed Compression

6.1 Introduction & Related Work

In this chapter, we introduce our uncertain distributed compression model, and describe and analyze our corresponding protocol.

Specifically, motivated in part by the goal of understanding human communication and in particular phenomena associated with the formation and development of language, we introduce a distributed compression problem and study it. We start with a description of the compression problem first, and then give our motivation.

6.1.1 Model

The Basic Model. We consider a distributed setting where $K$ players, with a complete network of point-to-point connections, are exchanging a sequence of messages drawn from an, apriori unknown, distribution $Q$. In our model, the set of possible messages is a countable set, and we use $\mathbb{N}$, the set of natural numbers to denote this set without loss of generality. The communication proceeds in rounds: In round $t$, a message $m$ is chosen from $\mathbb{N}$ according to $Q$ independently of the past. Simultaneously, an ordered pair of players $i, j \in [K] \overset{\text{def}}{=} \{1, \ldots, K\}$ with $i \neq j$ is chosen uniformly from all such pairs. The goal is for player $i$ to encode the message $m$ into a sequence of bits and send it to player $j$. Player $j$ receives this sequence of bits and decodes it to a message $\hat{m}$. (Note that the encoding, and decoding, may depend on the history of interactions involving the sender and receiver respectively.) Then, round $t$ is said to have an error if $m \neq \hat{m}$. The goal is to design encoding and decoding schemes that satisfy the condition that for every round $t$, the probability of error, over the history of random choices, is at most $\epsilon$ and the measure of performance is the expected length of communication averaged over the rounds up to $t$, studied as a function of $t$.

Efficiency Issues. A second measure of performance of the encoding and decoding algorithms is their “computational efficiency”. We define this notion using a “data-structural” perspective. Note that any encoder or decoder essentially needs to learn and store (approx-
imations to) the distribution $Q$ in order to perform moderately well. Thus, such an encoder or decoder needs to work with the amount of space that it might take to remember $Q$. At the same time, encoding or decoding a single message should not, and need not, take time linear in the storage. We thus measure the efficiency of the encoding and decoding algorithms in terms of its space requirement, and its processing time to compute the encoding of a message $m$ including the time it takes to update its memory to incorporate this new message in its history.

**Setup Assumptions.** Finally, we parametrize one commonality in the initialization of the different players. Note that to initialize any communication, the players must have some way of exchanging messages. One may consider the natural binary description of messages as one such possibility. Other possibilities may go via Kolmogorov complexity, i.e., by letting the players share a common universal machine and then representing a message via the encoding of the machine that outputs the binary representation of the message and halts. Rather than choosing any one of these representations, we parametrize the setup by the exact initial representation. More precisely, we consider a distribution $P$ on $\mathbb{N}$ for which a given initial representation is optimal and assume that all players share $P$ at the outset. Thus, the encoding and decoding algorithms may depend on this prior distribution, but otherwise the algorithms must be completely uniform and may not rely on any other shared information. Note that $Q$ is chosen adversarially and has no relationship to $P$. However, we will allow our main performance measure, i.e., the expected length of the compression to depend on the gap (distance or divergence) between $P$ and $Q$.

### 6.1.2 A Motivation: Language Formation and Development

One of our motivations to study this distributed compression problem is to give a fresh perspective on phenomena associated with the formation and evolution of human languages. We note that the study of languages is a central quest in linguistics, cognitive science and philosophy and much is known about it based on empirical studies. Our hope is to add some mathematical flavors to it.

For our purpose, we may view language as providing a map that describes how to convert a message in an individual’s brain into a sequence of utterances. Yet no language has a short description of this map. Part of the challenge seems to be that language is constantly evolving and if one were to fix any bound, language seems to evolve to a point where the description length exceeds this bound. The reason for this evolution may be viewed as some form of compression. While the ultimate goal may not be the time it takes to convey a message, language certainly evolves by creating shortcuts for currently frequent messages. This motivates our use of the compression capability of the message-to-utterance map as a

---

1Intuitively, we can think of $P$ as being a primitive “gesturing language” that is understandable to all people.

2For instance, a language could evolve to use the word “Ix” to denote “a boy who is not able to satisfactorily explain what a Hrung is, nor why it should choose to collapse on Betelgeuse Seven” [Ada79].
crude measure of performance. It is not the unique goal, but it is well-aligned with the goals of language.

A second feature of languages is that no two individuals probably have identical descriptions of the map. Attempts to give a unified description of the language (say, as in a dictionary) end up with many different dictionaries and each one capturing some segment of the population. Yet, language is robust to this variation and for the most part, communication manages to work despite the lack of agreement on the dictionary. Our view of this diversity is to consider the process of language acquisition. Individuals (children) learn from examples and indeed there is major diversity in the set of examples one encounters depending on one’s own circumstances, but even if one were to factor out this diversity (e.g., by considering identical twins), their experiences are still different. This inspires our setting: individuals are all born identical and get samples from the same distribution. (Furthermore, there are no network effects - the underlying graph is a complete graph and the message distribution is independent of the edge distribution. We will discuss this shortly.) Yet their samples are not identical and even this minor discrepancy seems to foil simple algorithms to coordinate on a compression map and introduces either diversity in the map, or complexity in the coordination process. Thus, the distributed compression problem already gives a potential reason for the diversity in language.

We emphasize that our choice of a simple graph (the complete graph) and the independence between the messages and the graph are not restrictions of the “model”. It is quite easy to extend our model to the setting where the graphs are complex, the distributions on edges are weighted and to allow the distribution of the messages to depend on the edge. While such richness is permitted by the model, we restrict to the simple setting to allow simpler contrasts between basic options (and our more sophisticated one).

Finally, one intriguing aspect of language is the amount of influence that different players have on its development. For the most part, language evolution seems to be a decentralized process, but this does not imply equal influence for all players. The role of books, especially those on grammar or dictionaries, of the media, and popular figures definitely assigns disproportionate influence to different players. A question that might be asked is whether language could manage to gain coherence across the population in the absence of such highly influential figures. Our model offers a way to study such questions (in our simplified setting of compression).

6.1.3 Context and Main Benchmarks

Our main result in this chapter is a distributed compression algorithm with “decent” performance. To set the stage for this algorithm, we first describe some basic benchmarks and then describe some basic compression schemes.

In what follows, we use \(Q(m)\) to denote the probability of a message \(m\) being drawn according to distribution \(Q\). We let \(H(Q) = \sum_{m \in \mathbb{N}} Q(m) \log_2(1/Q(m))\) denote the binary entropy of \(Q\) and we let \(D(Q||P) = \sum_{m \in \mathbb{N}} Q(m) \log_2(Q(m)/P(m))\) denote the KL-divergence between \(Q\) and \(P\). The best possible compression scheme would need at least \(H(Q)\) bits per message in expectation (even if all parties know \(Q\)) — this is true in the two-party case and
we will discuss below whether this is achievable in the distributed setting.

We refer to \( \tau = 2t/K \) as the local time, which roughly measures the number of messages any one player has seen (either as sender or receiver) at time \( t \). We use \( T_\epsilon \) to denote the local time by which a fixed player can obtain an \( \epsilon \)-close approximation to \( Q \), with probability at least \( 1 - \epsilon \). Note that \( T_\epsilon \) can be upper bounded by \( O(2^{H(Q)/\epsilon}) \) (and so in particular \( T_\epsilon \) is finite for distributions with finite entropy). Intuitively, \( T_\epsilon \) is a reasonable measure of local time by which one may expect to be able to compress well according to \( Q \) (even in the simple two-player setting) and this will also be a benchmark time for our compression algorithms.

Finally, a natural upper bound on the space complexity of storing (an \( \epsilon \)-approximation of) \( Q \) is again \( 2^{H(Q)/\epsilon} \). We will compare the storage needs of the various solutions below to this benchmark. Natural measures of update times would be polylogarithmic in space and we will ask for that. (In what follows, we assume messages are given as black boxes that can be stored in unit time and space and that basic operations such as comparison of messages (is \( m_1 \leq m_2 \)) take unit time.)

We now turn to some basic schemes for compression.

Near-Ideal Compression We first point out the (obvious?) flaw with the most natural hope one may have: Players could try to learn \( Q \) and get \( \epsilon \)-close to the right distribution moderately fast (in local time \( T_\epsilon \)) and then use the optimal (Huffman) code applied to such a distribution. Unfortunately, they cannot agree on this naive distribution and so no naive variation of the two-player compression mechanism seems to be implementable.

Static Compression: Players simply encode and decode according to the Huffman code for distribution \( P \). The error probability is zero and the expected length of the compression will be at most \( H(Q) + D(Q||P) + O(1) \). The good news with this scheme is that the performance does not depend on \( K \), but the bad news is that players do not learn to speak more effectively from examples. This is captured by the fact that the gap from optimal compression is \( D(Q||P) \) and we think of this as a large quantity.

Point-to-Point Compression: For every ordered pair \( (i, j) \), player \( i \) uses the Lempel-Ziv (or any universal) compression algorithm restricted to the sequence of messages that were directed from \( i \) to \( j \) and player \( j \) decodes according to the same history. This scheme converges to a compression length of \( H(Q) \) but it takes a relatively long time - a local time of \( K \times T_\epsilon \). We view dependence on \( K \) in the local time as too high. This scheme also involves memory requirement which is \( K \) times larger than the space needed for a two-player solution.

Dictatorial Compression: Here one player (the dictator) is singled out and tasked with the compression task. He learns a distribution close to \( Q \) and then communicates the resulting encoding/decoding scheme to all other players. The compression achieved by this scheme is near-optimal (converges to \( H(Q) \)); and the space requirement is also near-optimal. The main quantitative weakness we see is a mild dependence on \( K \) in the time it takes for this scheme to converge: Specifically it takes about \( T_\epsilon \) local time.
for the dictator to learn the distribution (which is perfectly fine), but then it needs to spread the information out to all $K$ players and this takes $T\epsilon + \Theta(\log K)$ additional local time (using any reasonable gossip algorithm with proper pipelining of messages). One main “criticism” of the scheme may be that it is centralized. While centralized mechanisms do plausibly play a role in the development of languages, they do not seem to be the only mechanism, and so we seek a truly distributed solution below.

6.1.4 Results

We now state our main theorem.

**Theorem 6.1.1 (Main Theorem).** Let $\epsilon > 0$ be a sufficiently small positive absolute constant. For all $K$ and $P$, there exists a deterministic distributed compression protocol $\Pi$ such that for any distribution $Q$ over $\mathbb{N}$, when run for $T$ iterations, it is the case that

- the amortized communication cost of $\Pi$ over $T$ iterations approaches $O(H(Q) + \log D(Q||P) + \log(1/\epsilon))$ as $T$ gets large. More precisely, the amortized communication cost is

$$O \left( H(Q) + \log D(Q||P) + \log(1/\epsilon) + \frac{2^{\Theta(H(Q) + D(Q||P)) / \epsilon} \cdot K \cdot D(Q||P) + 1}{T} \right).$$

- in each round, the transmitter and receiver run in time linear in their input and output sizes.

- the space usage is exponential in $(H(Q) + D(Q||P))/\epsilon$.

Our scheme is obtained with each player mixing the static scheme (used initially) with a switch to a more complex scheme once a sufficiently good approximation of $Q$ has been learned (by the player). A central ingredient in our scheme is a solution to the “Uncertain Compression” problem studied by Juba et al. [JKKS11] and Haramaty and Sudan [HS14b]. In the uncertain compression problem, two players attempt to compress a single message drawn from a distribution $Q$, but only the sender knows $P$ and the receiver only knows some distribution $Q'$ which is close to $Q$. The uncertain compression problem seems to arise naturally in our setting (neither the sender nor the receiver know $Q$ in our case, but both are close and this mild difference can simply be ignored). [JKKS11] give a “randomized” solution to this problem which compresses messages roughly down to $H(Q) + O(1)$ bits. Adapting this solution to our setting, essentially as a black box, achieves similar effects in our setting (compression down to $H(Q) + \delta \cdot D(Q||P)$ bits in time $\delta^{-1} \cdot T\epsilon$ local time), but a flaw with this scheme is that it requires the players to share a large random string in the setup phase. Instead, we turn to the solution of [HS14b] which does not need any randomness, but their solution assumes that $Q$ is supported on a finite set (of size $N$) and their compression length is $O(H(Q) + \log \log N)$. Since our distributions are not supported on finite sets, we need to modify their scheme and a careful modification followed by a relatively straightforward analysis leads to our eventual scheme and analysis. In the process, we are also able to build

167
a small data structure implementing the encoding and decoding with efficient processing time. We point out that if one does not care about computational efficiency, then we can remove the additive \( \log(1/\epsilon) \) term from the communication cost in Theorem 6.1.1 above while also replacing the multiplicative \( 2^{\Theta(H(Q) + D(Q||P))/\epsilon} \) factor by \( 2^{\Theta(H(Q))/\epsilon} \) (for more details, see Section 6.4).

### 6.1.5 Previous Work on Language Evolution

There have been many works on language evolution (to the best of our knowledge all from outside the theoretical computer science community). Without trying to be exhaustive, we briefly mention some of them. In the linguistics field, significant work has been done in the last decades on trying to understand language evolution, including [Cho75, Cho80]. Several papers also study language from the landscape of evolutionary game theory and evolutionary biology, e.g., [PB90, NK99, NPK99, Now00, NK01, NKN02, HCF02, KDG07, LMJ+07], and neuroscience, e.g., [RA98]. There has also been some previous attempts to connect language evolution to the framework of information theory (e.g., [PN00]), but their focus is on word formation in the “two-player” case, unlike our setup where we consider language as the outcome of the interaction between several players. To the best of our knowledge, the distributed compression perspective developed in this chapter has not been considered before.

### Organization

In Section 6.2, we formally define our distributed compression model. In Section 6.3, we describe our main protocol along with its computationally efficient implementation (Section 6.3.1, Section 6.3.2, Section 6.3.3, Section 6.3.4 and Section 6.3.5). In Section 6.4, we describe a computationally inefficient variant of our protocol that requires smaller communication. We conclude this chapter with some interesting open questions and future directions in Section 6.5.

### 6.2 Formal Definitions

Throughout this chapter, we denote by \( H(Q) \triangleq \sum_x Q(x) \log(1/Q(x)) \) the Shannon entropy of a probability distribution \( Q \), and by \( D(Q||P) \triangleq \sum_x Q(x) \log(Q(x)/P(x)) \) the KL divergence between probability distributions \( Q \) and \( P \). For any set \( S \) of elements, we write \( u \in_R S \) to mean that \( u \) is sampled uniformly at random from the set \( S \). We also denote by \( \mathbb{N} \) the set of all natural numbers.

We now formally define our uncertain distributed compression setup.

**Definition 6.2.1 (Distributed Compression).** A distributed compression protocol \( \Pi \) is parametrized by a tuple \((K, P, \epsilon)\) where

- \( K \) is the number of players.
• $P$ is a prior distribution over $\mathbb{N}$, which the players all agree on.

• $\epsilon$ is an error parameter.

The protocol is run on an instance parametrized by a pair $(Q, T)$ where $Q$ is the “true” distribution over $\mathbb{N}$, and $T$ is the total number of iterations for which the protocol is run. Both $Q$ and $T$ are unknown to the players. In any iteration $t \in [T]$:

• Two distinct players $i$ and $j$ are chosen uniformly at random from $[K]$.

• A message $m$ is sampled from distribution $Q$, and is given to player $i$.

• Player $i$ attempts to communicate $m$ to player $j$ by sending a single message comprising of $C_t$ bits.

• Player $j$ outputs a message $\hat{m}$.

The protocol is required to be such that, for any $Q$, and in any iteration $t$, it holds that $\Pr[\hat{m} \neq m] \leq \epsilon$, where the probability is over the randomness of the messages and players chosen in the history of the protocol. The amortized communication cost of $\Pi$ is defined to be $\sum_{t \in [T]} C_t/T$.

During the description of the protocol and the analysis, we will use $t$ to denote the current iteration. Also, we will use $t_i$ to denote the “local time” of player $i$. That is, $t_i$ is the number of times player $i$ was picked as the sender. Note that $t = \sum_{i \in [K]} t_i$.

6.3 A Distributed Compression Protocol

In this section, we prove the following theorem, which is the same as Theorem 6.1.1 but “without the computational efficiency” part. The proof of the computational efficiency part of Theorem 6.1.1 appears in Section 6.3.5.

**Theorem 6.3.1.** Let $\epsilon > 0$ be a sufficiently small positive absolute constant. For all $K$ and $P$, there exists a deterministic distributed compression protocol $\Pi$, such that for any distribution $Q$ over $\mathbb{N}$ when run for $T$ iterations, the amortized communication cost of $\Pi$ over $T$ iterations approaches $O(H(Q) + \log D(Q||P) + \log(1/\epsilon))$ as $T$ gets large.

More formally, for $T \geq 8 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \cdot K$, the amortized communication cost is

$$O \left( H(Q) + \log D(Q||P) + \log(1/\epsilon) + \frac{2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \cdot K}{T} \cdot D(Q||P) + 1 \right).$$

In the rest of this section, we describe the protocol behind the proof of Theorem 6.3.1.
6.3.1 Overview of the Protocol

We begin by giving a brief overview of the protocol. In any iteration, the chosen players will use one of two protocols that we call Static protocol (Section 6.3.2) and Uncertain protocol (Section 6.3.3).

On a high level, the Static protocol communicates messages with zero error, but it uses $H(Q) + D(Q||P) + O(1)$ bits of communication in expectation. On the other hand, the Uncertain protocol communicates $O(H(Q) + \log(D(Q||P)))$ bits in expectation, but it makes errors with some probability.

Suppose during iteration $t$, a message $m$ is chosen to be sent by player $i$ to player $j$, where $m$ is sampled from the unknown distribution $Q$. In this case, player $i$ will decide to communicate using either the Static protocol or the Uncertain protocol. Intuitively, in the initial few rounds in which player $i$ is the sender, she will use the Static protocol as she does not want to risk incurring large error by using the Uncertain protocol. But, once player $i$ has seen enough messages, she will switch to using the Uncertain protocol. The final bound on the amortized communication cost comes about by showing that the protocol ends up using the Uncertain protocol much more often than the Static protocol.

In Section 6.3.4, we describe exactly how the players switch between the two protocols and prove Theorem 6.3.1.

6.3.2 The Static Protocol

In the Static protocol, player $i$ uses the Huffman codebook for distribution $P$ in order to communicate the message $m$. The expected communication cost of doing so is $H(Q) + D(Q||P) + O(1)$. The nice aspects of this protocol are that the error probability is zero, and the players do not require any knowledge about the unknown distribution $Q$. However, the downside is that the communication cost is quite high in terms of the dependence on $D(Q||P)$.

We summarize the performance of the Static protocol in the following straightforward lemma.

**Lemma 6.3.2 (Static Protocol).** Suppose that during iteration $t$, a message $m$ is chosen to be sent by player $i$ to player $j$, where $m$ is sampled according to the unknown distribution $Q$. Then, player $i$ can communicate $m$ to player $j$ with zero error, such that the expected communication length is upper bounded by

$$H(Q) + D(Q||P) + 1.$$ 

6.3.3 The Uncertain Protocol

The Uncertain protocol is suitable when the players have individually learned good estimates of the distribution $Q$. However, since the players do not exactly agree on their learned estimates, we need an approach for the players to communicate when their estimates of $Q$
are close but may not be exactly identical. Our approach is inspired by [HS14b], and we obtain a protocol that in expectation communicates roughly $O(H(Q) + \log D(Q\|P)) + O(1)$ bits. We summarize the Uncertain protocol in the following lemma.

**Lemma 6.3.3 (Uncertain Protocol).** Suppose that during iteration $t$, a message $m$ is chosen to be sent by player $i$ to player $j$, where $m$ is sampled according to the unknown distribution $Q$. Then, player $i$ can communicate $m$ to player $j$, such that the expected communication length is upper-bounded by

$$O(H(Q) + \log D(Q\|P) + \log(1/\epsilon) + 1).$$

Moreover, the error probability is at most

$$2 \cdot e^{-\frac{\epsilon}{8}} + \frac{\epsilon}{4},$$

where the randomness is over all past messages and players chosen in the previous iterations.

**Isolating Hash Families** In order to describe the Uncertain protocol achieving Lemma 6.3.3, we will need the following notion of an isolating hash family which generalizes that of [HS14b].

**Definition 6.3.4 (Isolating Hash Families).** Let $N$, $R$ and $\ell$ be positive integers and $\epsilon \in (0,1]$. Then, a collection $H = \{h_1, h_2, \ldots, h_M : [N] \to [R]\}$ is said to be $(N, \ell, \epsilon)$-isolating if for every subset $S \subseteq [N]$ with $|S| \leq 2^\ell - 1$ and every $m \in [N] \setminus S$, we have that $\Pr_{h \in H}[h(m) \in h(S)] < \epsilon$. We call $M$ the size and $R$ the range-size of the isolating hash family $H$. The family $H$ is said to be efficiently computable if there is an algorithm that takes as input $i \in [M]$ and $j \in [N]$ and computes $h_i(j)$ in time polynomial in $\log M$, $\log N$ and $\log R$.

We note that the family used in [HS14b] corresponds to setting $\epsilon = 1$ in Definition 6.3.4. The next lemma shows the existence of an explicit and efficiently computable $(N, \ell, \epsilon)$-isolating hash family of relatively small size and small range-size.

**Lemma 6.3.5.** For every positive integers $N$ and $\ell$ and every $\epsilon \in (0,1]$, there exists an explicit and efficiently computable $(N, \ell, \epsilon)$-isolating hash family $H_{(N, \ell, \epsilon)}$ of size and range-size at most $2^\ell \cdot \frac{\log N}{\epsilon}$.

**Proof.** Let $q = 2^\ell \cdot [\log n + \log \frac{1}{\epsilon}]$. For each $x \in \mathbb{F}_q$, define the function $h_x$ to be the evaluation of the polynomial defined by $m$ on $x$, i.e.,

$$h_x(m_0, \ldots, m_{n-1}) \triangleq \sum_{i=0}^{n-1} m_i x^i.$$

By the fundamental theorem of algebra, for every $m' \neq m$, we have that $\Pr_x[h_x(m) = h_x(m')] \leq \frac{n}{q} \leq 2^{-\ell - \log \frac{1}{\epsilon}}$. Thus, by the union bound, for every set $S$ of size at most $2^\ell - 1$, we have that $\Pr[f(m) \in f(S)] \leq \epsilon$, as required. \hfill \Box
Pre-Processing Step  As stated earlier, all the players come in with a prior distribution \( P \). In addition, as part of the pre-processing, they compute and store the following:

- Divide the input space \( \mathbb{N} \) into a countable number of buckets indexed by \( r \in \mathbb{N}_{>0} \), given by \( A_r = \{ m : 2^{-r} < P(m) \leq 2^{-r+1} \} \). Clearly, for any \( r \), it holds that \( |A_r| \leq 2^r \).
  In addition, define the function \( r(m) := \lceil \log(1/P(m)) \rceil \) for every \( m \in \mathbb{N} \), that is, \( r(m) \) is the index of the bucket to which \( m \) belongs.

- For every \( r \), fix an (arbitrary) choice of isolating hash families \( \mathcal{H}_{(N,\ell,\epsilon/4)} \), for \( N = |A_r| \) and every choice of \( \ell \in \{1, 2, \ldots, \lceil \log N \rceil \} \).

Suppose that during iteration \( t \), a message \( m \) is chosen to be sent by player \( i \) to player \( j \), where \( m \) is sampled according to the unknown distribution \( Q \). Define \( Q_i^t \) to be the empirical distribution of the samples seen by player \( i \) up to iteration \( t \) (which includes the iteration \( t \), where the message seen is \( m \)). Similarly, define \( Q_j^t \) to be the empirical distribution of the samples seen by player \( j \) up to iteration \( t \) (this includes iteration \( t \), but by definition player \( j \) does not see any message in this iteration). The players use the encoding and decoding strategies described next.

Encoding  Upon receiving message \( m \), player \( i \) does the following:

(i) Let \( A \overset{\text{def}}{=} A_{r(m)} \) and \( N \overset{\text{def}}{=} |A| \).

(ii) Let \( \ell = \lceil \log(4/Q_i^t(m)) \rceil \).

(iii) Let \( u \in_R [\mathcal{H}_{(N,\ell,\epsilon/4)}] \).

(iv) Send the tuple \((r, \ell, u, h_u(m))\) to player \( j \).

The intuition for this encoding is as follows: upon receiving \( r \), player \( j \) understands that \( m \in A_r \), upon receiving \( \ell \), she understands which hash family to use, upon receiving \( u \), she knows which hash function to use, and hopefully with \( h_u(m) \), she will be able to recover \( m \) correctly.

Decoding  Upon receiving the tuple \((r, \ell, u, h_u)\), player \( j \) does the following:

(i) Set \( A = A_r \) and \( N \overset{\text{def}}{=} |A| \).

(ii) Identify \( h_u \in \mathcal{H}_{(N,\ell,\epsilon/4)} \).

(iii) Output \( \arg\max_{m' \in A : h_u(m') = h_u} Q_j^t(m') \).

Analysis  

We now analyze the operation of the above protocol.
**Communication Cost** Suppose the message \( m \) is chosen to be sent by player \( i \) to player \( j \). The communication cost of sending the tuple \((r, \ell, u, h_u(m))\) is as follows:

(i) \( \log [\log(1/P(m))] \) bits to send \( r \).

(ii) \( \log(\log(1/Q_i^i(m)) + 3) \) bits to send \( \ell \), since \( \ell \leq \log |S| + 1 \leq \log(4/Q_i^i(m)) + 1 \).

(iii) \( \log(1/Q_i^i(m)) + \log[\log(1/P(m))] + \log(1/\epsilon) + 5 \) bits to send \( u \) (it takes \( \ell + \log \log N + \log(4/\epsilon) \) bits).

(iv) \( \log(1/Q_i^i(m)) + \log[\log(1/P(m))] + \log(1/\epsilon) + 5 \) bits to send \( h_u(m) \).

Thus, the total communication is given by

\[
2 \log(1/Q_i^i(m)) + 3 \log[\log(1/P(m))] + \log(\log(1/Q_i^i(m)) + 3) + 10 + 2 \log(1/\epsilon).
\]

We wish to prove guarantees on the expected communication cost when \( Q \) is drawn from \( Q \). The terms in (III) are lesser order terms, which are smaller than (I), and thus we can ignore them. Term (II) in expectation is

\[
\mathbb{E}_{m \sim Q} \left[ \log \left( \left\lfloor \frac{1}{P(m)} \right\rfloor \right) \right] \leq \log \left( \mathbb{E}_{m \sim Q} \left[ \log \frac{1}{P(m)} \right] \right) \leq \log(H(Q) + D(Q||P) + 1).
\]

Term (I) is slightly more tricky to bound in expectation. Note that the empirical distribution changes on receiving message \( m \) (this turns out to be critical in bounding the communication!). That is, \( Q_i^i(m) = \frac{1+ t_{i-1}-Q_i^i(m)}{t_i} \). Also let \( \mathcal{M}_i^i \) be the multi-set of all messages that player \( i \) has seen up to time \( t \). Thus, Term (I) in expectation is as follows:

\[
\mathbb{E}_{\mathcal{M}_i^i} \mathbb{E}_{m \sim Q} \left[ \log \frac{1}{Q_i^i(m)} \right] = H(Q) + \mathbb{E}_{\mathcal{M}_i^i} \mathbb{E}_{m \sim Q} \left[ \log \frac{Q(m)}{\frac{1}{t_i} + \frac{1}{t_i} + \frac{1}{t_i}Q_i^i(m)} \right].
\]

In order to bound the second term above, we consider two cases: (i) \( Q_i^i(t-1)(m) \geq Q(m)/2 \) and (ii) \( Q_i^i(t-1)(m) < Q(m)/2 \). After fixing \( t_i \) and \( m \), by a Chernoff bound over the randomness of \( \mathcal{M}_i^i(t-1) \), we have that case (i) happens with probability at least \( 1 - \exp(-t_i \cdot Q(m)/8) \).

Moreover, we have that:

- **Case (i):** \( Q_i^i(t-1)(m) \geq Q(m)/2 \) \( \Rightarrow \) \( \log \left( \frac{Q(m)}{\frac{1}{t_i} + \frac{1}{t_i} Q_i^i(m)} \right) \leq 1 \).

- **Case (ii):** \( Q_i^i(t-1)(m) < Q(m)/2 \) \( \Rightarrow \) \( \log \left( \frac{Q(m)}{\frac{1}{t_i} + \frac{1}{t_i} Q_i^i(m)} \right) \leq \log(t_i \cdot Q(m)) \).
Using these upper bounds, we get that:

\[
\mathbb{E}_{\mathcal{M}^t_{(t-1)}} \mathbb{E}_{m \sim Q}\left[ \log \frac{Q(m)}{\frac{1}{t_i} + \frac{1}{t_i}Q_{(t-1)}^{i}(m)} \right] \leq \mathbb{E}_{m \sim Q} \left[ 1 \cdot (1 - e^{-t_i \cdot Q(m)/8}) + \log(t_i \cdot Q(m)) \cdot e^{-t_i \cdot Q(m)/8} \right] \\
\leq 1 + \mathbb{E}_{m \sim Q} \left[ \log(t_i \cdot Q(m)) \cdot e^{-t_i \cdot Q(m)/8} \right] \\
\leq 2,
\]

where the last inequality just follows from the fact that \( \log(x) \cdot e^{-x/8} \leq 1 \) for all \( x \).

Thus the overall communication is bounded by

\[
(2 + o(1)) H(Q) + 3 \log D(Q||P) + 2 \log(1/\epsilon) + O(1)
\]

**Error Guarantee** We now show that the error probability in iteration \( t \), denoted by \( p^t_{\text{err}} \) of the protocol is upper bounded by \( 2 \cdot e^{-\frac{1}{8} \frac{Q(m)}{K} t} + \epsilon/4 \), where \( m \) is fixed to be the message sent in round \( t \).

Since player \( i \) has communicated \((r, \ell, u)\), player \( j \) knows the correct bucket \( A_r \) of messages to which \( m \) belongs. Knowing \( \ell \) and \( u \), player \( j \) also knows which hash function is being used, which is chosen to ensure that for every set \( S \) of size \( \leq 2^\ell \), with probability \( 1 - \epsilon/4 \), for all \( m' \in S \setminus \{m\} \), \( h_u(m) \neq h_u(m') \).

Thus, if \( \ell \leq \log(1/Q^j_l(m)) \), then the \( j \)-th player will distinguish \( m \) from the set \( S = \{m' \in A \mid Q^j_l(m') \geq Q^j_l(m)\} \) with probability \( 1 - \epsilon/4 \). We now bound the probability that this does not happen:

\[
p^t_{\text{err}} \leq \Pr \left[ \ell > \log(1/Q^j_l(m)) \right] + \frac{\epsilon}{4} \\
\leq \Pr \left[ 4 \cdot Q^j_l(m) \geq Q^j_l(m) \right] + \frac{\epsilon}{4} \\
\leq \Pr \left[ Q^j_l(m) \geq 2 \cdot Q^j_l(m) \right] + \Pr \left[ Q^j_l(m) \leq \frac{1}{2} Q^j_l(m) \right] + \frac{\epsilon}{4} \\
\leq e^{-\frac{1}{8} \frac{Q^j_l(m)}{K} t} + e^{-\frac{1}{8} \frac{Q^j_l(m)}{K} t} + \frac{\epsilon}{4},
\]

where the last inequality follows from the Chernoff bound and the fact that \( Q^j_l \) and \( Q^j_l \) are binomial distributions with parameters \( t \) and \( \frac{Q^j_l(m)}{K} \).

### 6.3.4 Final Protocol

We are now ready to present the protocol desired in Theorem 6.3.1. As before, suppose that during iteration \( t \), a message \( m \) is chosen to be sent by player \( i \) to player \( j \), where \( m \) is sampled according to the unknown distribution \( Q \). As defined in Section 6.3.3, define \( Q^j_l \) to be the empirical distribution of the samples seen by player \( i \) up to iteration \( t \) (which includes the iteration \( t \), where the message seen is \( m \)). Similarly, define \( Q^j_l \) to be the empirical
distribution of the samples seen by player $j$ up to iteration $t$ (this includes iteration $t$, but by definition player $j$ does not see any message in this iteration).

For ease of presentation, we will first assume that the players know the entropy of the distribution $Q$. This is not a natural assumption, and we get around it in Section 6.3.4. However, we first will describe the main protocol with this assumption to make the analysis more intuitive.

**Encoding**  Upon receiving message $m$, player $i$ does the following:

- If $t_i < 80 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon)$,
  - send the bit $b = 0$
  - use the STATIC protocol (Lemma 6.3.2) to send message $m$.

- Else,
  - send the bit $b = 1$
  - use the UNCERTAIN protocol (Lemma 6.3.3) to send message $m$.

(Here, the bit $b$ indicates whether player $i$ is using the STATIC protocol or the UNCERTAIN protocol).

**Decoding**  Depending on the value of the received bit $b$, player $j$ uses either the STATIC protocol or the UNCERTAIN protocol to decode and output $\hat{m}$.

**Analysis**

We now upper-bound the amortized communication cost and the error probability in any iteration of the above protocol.

**Communication Cost**  By the design of the final protocol, each player uses the STATIC protocol at most $80 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon)$ times, and hence overall, the STATIC protocol is used at most $O(2^{2H(Q)/\epsilon} \cdot \log(1/\epsilon) \cdot K)$ times. Thus, if the total number of iterations is $T$, then the total communication cost in expectation is at most

$$O \left( 2^{2H(Q)/\epsilon} \cdot \log(1/\epsilon) \cdot K \cdot (H(Q) + D(Q||P) + 1) + T \cdot O \left( H(Q) + \log D(Q||P) + O(1) \right) \right).$$

And hence, the expected amortized communication cost is at most

$$O \left( H(Q) + \log D(Q||P) + \frac{2^{2H(Q)/\epsilon} \cdot \log(1/\epsilon) \cdot K}{T} \cdot D(Q||P) + 1 \right).$$

175
Error Guarantee We first show the following lemma, which is an easy consequence of Markov’s inequality.

**Lemma 6.3.6.** For any distribution $Q$ over $\mathbb{N}$, it holds that,

$$
\Pr_{m \sim Q} \left[ Q(m) \geq 2^{-H(Q)/\epsilon} \right] \geq 1 - \epsilon.
$$

**Proof.** By the definition of the entropy $H(Q)$, we have that $\mathbb{E}_{m \sim Q} \left[ \log \frac{1}{Q(m)} \right] = H(Q)$. Thus, the following application of Markov’s inequality immediately implies the lemma:

$$
\Pr_{m \sim Q} \left[ \log \frac{1}{Q(m)} \geq \frac{H(Q)}{\epsilon} \right] \leq \epsilon.
$$

We will show that in any iteration $t$, the error probability is at most $\epsilon$, where the randomness is over all the past and current messages and chosen players. We distinguish two cases:

**Case 1.** If $t < 8 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \cdot K$:

Using the Chernoff bound, it is easy to see that

$$
\Pr \left[ t_i > 80 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \mid t < 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \cdot K \right] \leq \exp \left[ -\Omega \left( 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \right) \right] \ll \epsilon.
$$

Thus, it follows that with probability $\geq 1 - \epsilon$, player $i$ uses the Static protocol in which case there is zero error. Thus, the probability of error is at most $\epsilon$.

**Case 2.** If $t \geq 8 \cdot 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon) \cdot K$:

Lemma 6.3.6 implies that when a message $m$ is sampled from $Q$, with probability at least $1 - \epsilon/2$ it holds that $Q(m) \geq 2^{-2H(Q)/\epsilon}$. In this situation, player $i$ may choose to use either the Static or the Uncertain protocol. In the former case, the protocol makes no error. In the latter case, by Lemma 6.3.3, the protocol makes error with probability at most

$$
2 \cdot e^{-\frac{1}{8} K \frac{1}{Q(m)}} + \epsilon \leq 2 \cdot e^{-\frac{1}{8} 2^{-2H(Q)/\epsilon} 2^{2H(Q)/\epsilon} \cdot \log(8/\epsilon)} + \epsilon,
$$

which is at most $\epsilon/2$ if $Q(m) \geq 2^{-2H(Q)/\epsilon}$. Hence, we get that the total error probability is at most $\epsilon$.

**Getting around the entropy assumption**

We let $\epsilon > 0$ be a sufficiently small positive absolute constant. We now informally describe how to construct a protocol that does not assume that the players know the entropy of the distribution $Q$. We note that the main reason for the “switching” criterion $t_i < 80 \cdot 2^{2H(Q)/\epsilon}$.
\[
\log(8/\epsilon)\] was to ensure that when we are using the UNCERTAIN protocol and we encounter a message \(m\) with \(Q(m) \geq 2^{-2H(Q)/\epsilon}\) (which happens with probability at least \(1 - \epsilon/2\)), it holds that \(t_i \cdot Q(m) \gg \log(1/\epsilon)\).

Thus, the protocol guarantees will still hold as long as the players switch to the UNCERTAIN protocol after a sufficiently “large” time \(t_i\). Indeed, we show that it is possible to switch to the UNCERTAIN protocol after time \(t_i\) such that \(\Pr_{m \sim Q}[t_i \cdot Q(m) \gg \log(1/\epsilon)] \geq 1 - \epsilon/4\).

We now describe the “switching” criterion. In what follows, we prove that for every player, the switching criterion is not met too early, nor is it met too late. Lemma 6.3.9 shows that the probability that the switching criterion is met “too early” (i.e., before the time \(T_0\) defined below) is very small. Moreover, it turns out that the probability that the switching criterion is met “too late” (i.e., after time \(2^{O\left(\frac{H(Q)}{\epsilon} \cdot K\right)}\)) is also very small (see Lemma 6.3.7 below). Together, these two properties allow individual players to switch from the Static protocol to the UNCERTAIN protocol based on their observed history of messages. In turn, this allows us to carry out an analysis of the communication cost and the error probability without knowledge of the entropy of \(Q\).

We say that at player \(i\), the switching criterion is met at iteration \(t_i\) if
\[
t_i \geq \epsilon^{-3} \quad \text{and} \quad \sum_{m:Q_i(m) > t_i^{-\frac{1}{2}}} Q_i^i(m) \geq 1 - \frac{\epsilon}{2}.
\]

We first show that, with high probability, the switching criterion is met in (global) time \(2^{O\left(\frac{H(Q)}{\epsilon} \cdot K\right)}\).

**Lemma 6.3.7.** For every player \(i\), the probability that the switching criterion is met before time \(t > 4 \cdot 2^{\frac{4H(Q)}{\epsilon} \cdot \frac{t}{K}}\) is at least \(1 - \exp\left(-\frac{1}{64}\epsilon^2 2^{-\frac{4H(Q)}{\epsilon} \cdot \frac{t}{K}}\right)\).

**Proof.** Let \(m\) be such that \(Q(m) \geq 2^{-\frac{4H(Q)}{\epsilon}}\). By the Chernoff bound,
\[
\Pr\left[Q_i^i(m) \leq (1 - \frac{\epsilon}{4})Q(m)\right] \leq \exp\left(-\frac{1}{32}\epsilon^2 Q(m)\frac{t}{K}\right).
\]

Moreover, by the Chernoff bound, we have that \(\Pr[t_i \leq \frac{t}{2K}] \leq \exp\left(-\frac{t}{8K}\right)\). We define the event
\[
E = \left[t_i \leq \frac{t}{2K} \lor \exists m: Q(m) \geq 2^{-\frac{4H(Q)}{\epsilon}} \land Q_i^i(m) \leq (1 - \frac{\epsilon}{4})Q(m)\right].
\]

By the union bound, we get that
\[
\Pr[E] \leq \exp\left(-\frac{t}{8K}\right) + \exp\left(\frac{4H(Q)}{\epsilon} - \frac{1}{32}\epsilon^2 2^{-\frac{4H(Q)}{\epsilon} \cdot \frac{t}{K}}\right)
\]
\[
\leq \exp\left(-\frac{1}{64}\epsilon^2 2^{-\frac{4H(Q)}{\epsilon} \cdot \frac{t}{K}}\right).
\]

177
If the event $E$ does not hold, then for every $m$ that satisfies $Q(m) > 2^{-\frac{4H(Q)}{e}}$, we get that

$$Q_i(m) > (1 - \frac{\epsilon}{4})Q(m) > (1 - \frac{\epsilon}{4})2^{-\frac{4H(Q)}{e}} > \left(\frac{t}{2K}\right)^{-\frac{1}{2}} > t_i^{-\frac{1}{2}}.$$

Thus,

$$\sum_{m:Q_i(m) > t_i^{-\frac{1}{2}}} Q_i(m) \geq \sum_{m:Q(m) > 2^{-\frac{4H(Q)}{e}}} Q_i(m) \geq (1 - \frac{\epsilon}{4}) \cdot \sum_{m:Q(m) > 2^{-\frac{4H(Q)}{e}}} Q(m) = (1 - \frac{\epsilon}{4}) \cdot \Pr_{m \sim Q} [Q(m) > 2^{-\frac{4H(Q)}{e}}] \geq (1 - \frac{\epsilon}{4}) \cdot (1 - \frac{\epsilon}{4}) \geq 1 - \frac{\epsilon}{2}.$$

Moreover, $t_i > \frac{t}{2K} > \epsilon^{-3}$. Hence, in this case, the switching criterion is met.

Let $T_0$ be the smallest $t > \frac{1}{\epsilon K}$ that satisfies $\Pr_{m \sim Q} [Q(m) \geq \frac{1}{\sqrt{K}}] > 1 - \frac{3}{4} \epsilon$. First, we will observe that after time $T_0$, it is indeed safe to switch to the Uncertain protocol.

**Observation 6.3.8.** For every time $t \geq T_0$, the c protocol succeeds with probability at least $1 - \epsilon$.

**Proof.** By Lemma 6.3.3, with probability at least $1 - \frac{3}{4} \epsilon$, the protocol succeeds with probability at least $1 - \frac{\epsilon}{4}$.

It remains to show that with high probability, we will not use the Uncertain protocol before $T_0$.

**Lemma 6.3.9.** The probability that player $i$ meets the switching criterion before time $T_0$ is at most $\epsilon$.

**Proof.** We will show that for any fixed $t_i \leq \frac{2T_0}{K}$, we have that the probability that for player $i$, the switching criterion is met in local time $t_i$, is at most $2 \cdot \exp \left( -\frac{1}{12} \sqrt{t_i} \right)$. By the union bound, we will get that the probability that for player $i$, the switching criterion is met before local time $\frac{2T_0}{K}$ is bounded by

$$\sum_{t_i = \epsilon^{-3} + 1}^{\infty} 2 \cdot \exp \left( -\frac{1}{12} \sqrt{t_i} \right) \leq \int_{\epsilon^{-3}}^{\infty} 2 \cdot \exp \left( -\frac{1}{12} \sqrt{t_i} \right) dt_i = 24 \cdot (\sqrt{\epsilon^{-3} + 12}) \cdot \exp \left( -\frac{1}{12} \sqrt{\epsilon^{-3}} \right) \leq \frac{\epsilon}{2}.$$ 

178
Moreover, by the Chernoff bound, we have that the probability that the local time of player \( i \) in (global) time \( T_0 \) exceeds \( \frac{2T_0}{K} \) is at most \( \exp(-\frac{T_0}{3K}) \). Thus, the probability that for player \( i \) the switching criterion is met before time \( T_0 \) is at most \( \frac{\epsilon}{2} + \exp(-\frac{T_0}{3K}) \leq \epsilon \), as required.

Fix \( t_i \) and let \( M = \left\{ m \in \mathbb{N} \mid Q(m) \geq \frac{1}{2\sqrt{t}_i} \right\} \). Since \( t_i < \frac{2T_0}{K} \), we have that

\[
\Pr_{m \sim Q}[m \in M] = \Pr_{m \sim Q}[Q(m) \geq \frac{1}{2\sqrt{t}_i}] \leq \Pr_{m \sim Q}[Q(m) \geq \sqrt{\frac{K}{2\sqrt{2T_0}}}] \leq \Pr_{m \sim Q}[Q(m) \geq \sqrt{\frac{K}{4\sqrt{T_0}}}] \leq 1 - \frac{3}{4} \epsilon.
\]

Thus, by the Chernoff bound,

\[
\Pr \left[ \sum_{m \in M} Q_i^t(m) > 1 - \frac{\epsilon}{2} \right] \leq \exp \left( -\frac{\epsilon \cdot t_i}{25} \right). \tag{6.1}
\]

Now we upper-bound \( \Pr \left[ \exists m \notin M : Q_i^t(m) > \frac{1}{2\sqrt{t}_i} \right] \). To prove this bound, we can assume without lost of generality that for all \( m \) except one, we have that \( Q(m) > \frac{1}{5\sqrt{t}_i} \): if there exist two elements of such a small probability, we can merge them together to a single element and only increase the probability \( \Pr \left[ \exists m \notin M : Q_i^t(m) > \frac{1}{2\sqrt{t}_i} \right] \). So we will assume that there are at most \( 5\sqrt{t}_i + 1 \) such elements. By the Chernoff bound, we have that for each \( m \notin M \), \( \Pr \left[ Q_i^t(m) > \frac{1}{\sqrt{t}_i} \right] \leq \exp \left( -\frac{1}{6} \sqrt{t}_i \right) \) and by a union bound we can get that

\[
\Pr \left[ \exists m \notin M : Q_i^t(m) > \frac{1}{\sqrt{t}_i} \right] \leq (5\sqrt{t}_i + 1) \exp \left( -\frac{1}{6} \sqrt{t}_i \right) \leq \exp \left( -\frac{1}{12} \sqrt{t}_i \right). \tag{6.2}
\]

By Combining Equations (6.1) and (6.2), assuming \( t_i \geq \epsilon^{-3} \), we get

\[
\Pr \left[ \sum_{m : Q_i^t(m) > \frac{1}{\sqrt{t}_i}} Q_i^t(m) \right] \leq \Pr \left[ \sum_{m \in M} Q_i^t(m) \geq 1 - \frac{\epsilon}{2} \vee \exists m \notin M : Q_i^t(m) > \frac{1}{\sqrt{t}_i} \right] \leq 2 \exp \left( -\frac{1}{12} \sqrt{t}_i \right).
\]

This gives an upper bound of \( 2 \cdot \exp \left( -\frac{1}{12} \sqrt{t}_i \right) \) on the probability that at player \( i \), the switching criterion is met in local time \( t_i \), as needed.

\[\square\]

### 6.3.5 Efficient Implementation

We briefly sketch how to efficiently implement the encoding and decoding strategies of Section 6.3. The details are deferred to the full version. The overall update time will be linear in \( (H(Q) + D(Q||P))/\epsilon \), and the used memory will be proportional to the dictionary-size which is exponential in \( (H(Q) + D(Q||P))/\epsilon \). The key question of interest is how to compute the uncertain compression function efficiently. Note that while we would like a
fast “processing time” per update, the model naturally allows us to amortize the cost over many operations. In particular, the switch from the Static protocol to the Uncertain one does not have to be carried out in an instant. We will exploit this feature strongly. The corresponding efficient algorithm will have three phases:

1. A phase where we simply use the Static protocol while updating the empirical distributions.

2. A phase where the encoding and decoding dictionaries are being built, but where we still use the Static protocol.

3. A phase where we use the Uncertain protocol.

In what follows, we assume that the messages $m$ and the prior distribution $P$ are presented jointly so that the message $m$ given to player $i$ in round $t$ is $E_P(m)$, namely the Static (Huffman) encoding of $m$ under $P$. This is a natural assumption about $P$ — after all $P$ is meant to represent a simple and natural, though unoptimized, distribution over the message space. We now recall the statement of Theorem 6.1.1.

**Theorem 6.3.10.** Let $\epsilon > 0$ be a sufficiently small positive absolute constant. For all $K$ and $P$, there exists a deterministic distributed compression protocol $\Pi$, such that for any distribution $Q$ over $\mathbb{N}$ when run for $T$ iterations,

- the amortized communication cost of $\Pi$ over $T$ iterations approaches $O(H(Q) + \log D(Q||P) + \log(1/\epsilon))$ as $T$ gets large. More formally, the amortized communication cost is

  $$O \left( H(Q) + \log D(Q||P) + \log(1/\epsilon) + \frac{2^{\Theta(H(Q) + D(Q||P))/\epsilon}}{T} \cdot K \cdot D(Q||P) + 1 \right).$$

- in each round, the transmitter and receiver run in time linear in their input and output sizes.

- the space usage is exponential in $(H(Q) + D(Q||P))/\epsilon$.

Note that in Theorem 6.1.1, the input to the transmitter is $E_P(m)$ and the input to the receiver is the message that she gets from the transmitter.

**Proof Sketch.** Let $T_\epsilon = 2^{\Theta(H(Q) + D(Q||P))/\epsilon}$ denote the local time at which our inefficient transmitter and receiver – described in the previous section – should switch from the Static protocol to the Uncertain one. In the efficient protocol, during the execution of the Static protocol for the first $T_\epsilon$ units of local time, each player will also maintain a count of the number of times she has seen each message using a simple binary tree indexed by $E_P(m)$. At local time $T_\epsilon$, player $i$ updates his empirical distribution $Q_{T_\epsilon}^i$. Note that we can amortize this update time over several rounds. After round $T_\epsilon$, the efficient protocol will start building an encoding and decoding table for the uncertain compression algorithm, but will take $T'' = \text{poly}(T_\epsilon)$ rounds to do so (as we will explain below), and in the meanwhile, it will
continue using the Static protocol for these $T'$ rounds. At round $T_e + T'$, it will then switch to the Uncertain protocol, and at this stage it will have a complete table (for all relevant messages) for the encoding and decoding functions, and so it can encode and decode by a simple table lookup.

We also note that the upper bound on the amortized communication cost follows from a similar argument as in the proof of Theorem 6.3.1 in Section 6.3.

So it suffices to show that the encoding and decoding tables can be computed in time $\text{poly}(T_e)$. A straightforward implementation of the algorithm used in the proof of Theorem 6.3.1 essentially works, with a few additional observations. First, we note that we do not need to encode messages $m$ with $P(m) \leq 2^{-\Theta(H(Q) + D(Q||P))/\epsilon}$ since by Markov’s inequality such messages occur with probability less than $\epsilon$. This makes sure that the hash families that we need work with a value of $N$ which is at most $2^{(H(Q) + D(Q||P))/\epsilon}$ and the log $N$ factor in the size of these hash families is equal to $(H(Q) + D(Q||P))/\epsilon$, which is affordable. Next, we use the efficiently computable hash functions which are given by Lemma 6.3.5. We apply these hash functions to $E_P(m)$ rather than $m$ in order to make sure that their domain is also small. The upper bound on the encoding time now follows.

For the decoding time, we note that filling in one entry of the decoding table takes time linear in $N$ which is exponentially larger than the budget in the statement of Theorem 6.1.1. However, we can divide this task over $N$ rounds while performing $O(1)$ computations per round. The upper bound on the decoding time now follows.

Finally, the space usage is proportional to the size of the encoding and decoding lookup tables which is exponential in $(H(Q) + D(Q||P))/\epsilon$.

\[ \square \]

6.4 A Computationally Inefficient Protocol with Smaller Communication

In this section, we show that if one does not need to ensure computational efficiency, then we can remove the additive $\log(1/\epsilon)$ term from the communication cost in Theorem 6.1.1 while also replacing the multiplicative $2^{\Theta(H(Q) + D(Q||P))/\epsilon}$ factor by $2^{\Theta(H(Q))/\epsilon}$.

The general structure of the protocol is similar to the one in Section 6.3 except that for the description and analysis of the Uncertain protocol (Section 6.3.3). We now describe a computationally inefficient variant of the Uncertain protocol which has smaller communication. The performance of this variant is summarized in the following lemma.

Lemma 6.4.1 (Uncertain Protocol). Suppose that during iteration $t$, a message $m$ is chosen to be sent by player $i$ to player $j$, where $m$ is sampled according to the unknown distribution $Q$. Then, player $i$ can communicate $m$ to player $j$, such that the expected communication length is upper-bounded by

\[ O(H(Q) + \log D(Q||P)) + O(1). \]
Moreover, the error probability is at most
\[
\frac{1}{Q(m)} \cdot \exp\left(-\Omega\left(\frac{t \cdot Q(m)}{K}\right)\right),
\]
where the randomness is over all past messages and players chosen in the previous iterations.

We now describe the corresponding encoding and decoding procedures (along with the pre-processing step). Recall Definition 6.3.4 of an \((N, \ell, \epsilon)\)-isolating hash family. We now define an \((N, \ell)\)-isolating hash family to be an \((N, \ell, 1)\)-isolating hash family.

**Pre-Processing Step**  As stated earlier, all the players come in with a prior distribution \(P\). In addition, as part of the pre-processing, they perform the following steps (and store the outcomes):

- Divide the input space \(\mathbb{N}\) into a countable number of buckets indexed by \(r \in \mathbb{N}_{>0}\), given by \(A_r = \{m : 2^{-r} < P(m) \leq 2^{-r+1}\}\). Clearly, for any \(r\), it holds that \(|A_r| \leq 2^r\).

  In addition, define the function \(r(m) := \lceil \log(1/P(m)) \rceil\) for every \(m \in \mathbb{N}\), that is, \(r(m)\) is the index of the bucket to which \(m\) belongs.

- For every \(r\), fix an (arbitrary) choice of isolating hash families \(\mathcal{H}_{(N, \ell)}\), for \(N = |A_r|\) and every choice of \(\ell \in \{1, 2, \ldots, \lceil \log N \rceil\}\).

Suppose that during iteration \(t\), a message \(m\) is chosen to be sent by player \(i\) to player \(j\), where \(m\) is sampled according to the unknown distribution \(Q\). Define \(Q_i^t\) to be the empirical distribution of the samples seen by player \(i\) up to iteration \(t\) (which includes the iteration \(t\), where the message seen is \(m\)). Similarly, define \(Q_j^t\) to be the empirical distribution of the samples seen by player \(j\) up to iteration \(t\) (this includes iteration \(t\), but by definition player \(j\) does not see any message in this iteration). The players use the encoding and decoding strategies described next.

**Encoding**  Upon receiving message \(m\), player \(i\) does the following:

(i) let \(A \overset{\text{def}}{=} A_{r(m)}\) and \(N \overset{\text{def}}{=} |A|\).

(ii) let \(S \overset{\text{def}}{=} \{m' \in A \setminus \{m\} : Q_i^t(m') \geq \frac{1}{16} Q_i^t(m)\}\).

(iii) let \(\ell = \lceil \log |S| \rceil\).

(iv) let \(u \in [|\mathcal{H}_{(N, \ell)}|]\) and \(h_u \in \mathcal{H}_{(N, \ell)}\) such that \(h_u(m) \notin h_u(S)\).

(v) Send the tuple \((r, \ell, u, h_u(m))\) to player \(j\).

Note that the property of isolating hash families (see Definition 6.3.4) guarantees the existence of \(h_u \in \mathcal{H}_{(N, \ell)}\) as desired in (iv).
The intuition for this encoding is as follows: upon receiving $r$, player $j$ understands that $m \in A_r$, upon receiving $\ell$, she understands which hash family to use, upon receiving $u$, she knows which hash function to use, and hopefully with $h_u(m)$, she will be able to recover $m$ correctly.

**Decoding** Upon receiving the tuple $(r, \ell, u, h_u)$, player $j$ does the following:

(i) Set $A = A_r$ and $N \overset{\text{def}}{=} |A|$.  
(ii) Identify $h_u \in \mathcal{H}(N, \ell)$.  
(iii) Output $\arg \max_{m' \in A : h_u(m') = h_u Q_j^t(m')}$.  

The analysis of the communication cost and the error guarantee appears in Subsection 6.4.1, where Lemma 6.4.1 is proved.

### 6.4.1 Analysis of Computationally Inefficient Protocol

We now analyze the operation of the computationally inefficient protocol described above.

**Communication Cost** Suppose the message $m$ is chosen to be sent by player $i$ to player $j$. The communication cost of sending the tuple $(r, \ell, u, h_u)$ is as follows:

(i) $\log \lceil \log (1/P(m)) \rceil$ bits to send $r$.  
(ii) $\log (1/Q_i^t(m)) + 5$ bits to send $\ell$, since $\ell \leq \log |S| + 1 \leq \log (16/Q_i^t(m)) + 1$.  
(iii) $\log (1/Q_i^t(m)) + \log \lceil \log (1/P(m)) \rceil + 5$ bits to send $u$ (it takes $\ell + \log N$ bits).  
(iv) $\log (1/Q_i^t(m)) + 5$ bits to send $h_u(m)$.  

Thus, the total communication is given by,

$$2 \log (1/Q_i^t(m)) + 2 \log \lceil \log (1/P(m)) \rceil + \log (1/Q_i^t(m)) + 5 + 10.$$  

We wish to prove guarantees on the expected communication cost, when $m$ is drawn from $Q$. The terms in (III) are lesser order terms, which are smaller than (I), and thus we choose to ignore them. Term (II) in expectation is

$$\mathbb{E}_{m \sim Q} \left[ \log \left( \log \frac{1}{P(m)} \right) \right] \leq \log \left( \mathbb{E}_{m \sim Q} \left[ \log \frac{1}{P(m)} \right] \right) \leq \log (H(Q) + D(Q \parallel P) + 1).$$  

Term (I) is slightly more tricky to bound in expectation. Note that the empirical distribution changes on receiving message $m$ (this turns out to be critical in bounding the communication!). That is, $Q_i^t(m) = \frac{1 + (t-1)Q_i^{t-1}(m)}{t}$. Also let $\mathcal{M}_i^t$ be the multi-set of all messages that
player $i$ has seen up to time $t$. Thus, Term (I) in expectation is as follows,

$$ \mathbb{E}_{M_{(t-1)}^i} \mathbb{E}_{m \sim Q} \left[ \log \frac{1}{Q_i^t(m)} \right] = H(Q) + \mathbb{E}_{M_{(t-1)}^i} \mathbb{E}_{m \sim Q} \left[ \log \frac{Q(m)}{\frac{1}{t_i} + \frac{(t_i-1)Q_i^{t-1}(m)}{t_i}} \right] $$

In order to bound the second term above, we consider two cases, (i) $Q_i^{t-1}(m) \geq Q(m)/2$ or (ii) $Q_i^{t-1}(m) < Q(m)/2$. After fixing $t_i$ and $m$, by Chernoff bound over the randomness of $M_{(t-1)}^i$ we have that case (i) happens with probability at least $1 - \exp(-t_i \cdot Q(m)/8)$.

- Case (i) $Q_i^{t-1}(m) \geq Q(m)/2 \implies \log \left( \frac{Q(m)}{\frac{1}{t_i} + \frac{(t_i-1)Q_i^{t-1}(m)}{t_i}} \right) \leq 1$
- Case (ii) $Q_i^{t-1}(m) < Q(m)/2 \implies \log \left( \frac{Q(m)}{\frac{1}{t_i} + \frac{(t_i-1)Q_i^{t-1}(m)}{t_i}} \right) \leq \log(t_i \cdot Q(m))$

Using these upper bounds, we get that

$$ \mathbb{E}_{M_{(t-1)}^i} \mathbb{E}_{m \sim Q} \left[ \log \frac{Q(m)}{\frac{1}{t_i} + \frac{(t_i-1)Q_i^{t-1}(m)}{t_i}} \right] \leq \mathbb{E}_{m \sim Q} \left[ 1 \cdot \left(1 - e^{-t_i \cdot Q(m)/8}\right) + \log(t_i \cdot Q(m)) \cdot e^{-t_i \cdot Q(m)/8}\right] $$

$$ \leq 1 + \mathbb{E}_{m \sim Q} \left[ \log(t_i \cdot Q(m)) \cdot e^{-t_i \cdot Q(m)/8}\right] $$

$$ \leq 2, $$

where the last inequality just follows from the fact that $\log(x) \cdot e^{-x/8} \leq 1$ for all $x$. Thus the overall communication is bounded by

$$ (2 + o(1)) H(Q) + 2 \log D(Q\|P) + O(1). $$

**Error Guarantee** We now show that the error probability in iteration $t$, denoted by $\rho_t^{\text{err}}$ of the protocol is upper bounded by $\frac{1}{Q_i(m)} \cdot 2^{-\Omega(t_i \cdot Q(m)/8)}$, where $m$ is fixed to be the message sent in round $t$. We first give an intuitive explanation for the error bound. Since player $i$ has communicated $(r, \ell, u)$, player $j$ knows the correct bucket of messages $A_r$ to which $m$ belongs. Knowing $\ell$ and $u$, player $j$ also knows which hash function is being used, which is chosen to ensure that for every $m' \in S \setminus \{m\}$, $h_u(m) \neq h_u(m')$. Thus, the only way in which an error can happen is that there exists some $m' \notin S$ such that $h_u(m) = h_u(m')$ and $Q_i^t(m') > Q_i^t(m)$. Since $m' \notin S$, it implies by definition of $S$ that $Q_i^t(m') \leq Q_i^t(m)/16$, which means that player $i$ has seen the message $m'$ significantly fewer times compared to the message $m$. On the other hand, we also have that $Q_i^t(m') > Q_i^t(m)$, which means that player $j$ has seen the message $m'$ at least as many times as message $m$. For “large” $t_i$, it is very unlikely that players $i$ and $j$ have seen $m$ and $m'$ in such disproportionate manner.

To make the arguments go through, we need to union bound over all $m' \in A_r \setminus S$. However, a naive union bound is too lossy because we do not have any reasonable upper bound on the number of $m'$s. To get around this issue, we do a simple bucketing argument.
The formal upper bound on $p_t^{\text{err}}$ is shown as follows,

$$p_t^{\text{err}} = \Pr[\exists m' \in A : h_u(m') = h_u(m) \text{ and } Q_i^{\text{err}} m' > Q_i^{\text{err}} m]\n\leq \Pr[\exists m' \in A : Q_i^{\text{err}} m' < \frac{1}{16} Q_i^{\text{err}} m \text{ and } Q_i^{\text{err}} m' > Q_i^{\text{err}} m]$$

$$\leq \Pr[\exists m' \in A : Q(m') > \frac{1}{4} Q(m) \text{ and } Q_i^{\text{err}} m' < \frac{1}{16} Q_i^{\text{err}} m]$$

$$+ \Pr[\exists m' \in A : Q(m') \leq \frac{1}{4} Q(m) \text{ and } Q_i^{\text{err}} m' > Q_i^{\text{err}} m]$$

$$\leq \Pr[\exists m' \in A : Q(m') > \frac{1}{4} Q(m) \text{ and } Q_{i-1}^{\text{err}} m' < \frac{1}{16} \left(Q_{i-1}^{\text{err}} m + \frac{1}{t_i - 1}\right)]$$

$$+ \Pr[\exists m' \in A : Q(m) \leq \frac{1}{4} Q(m) \text{ and } Q_i^{\text{err}} m' > Q_i^{\text{err}} m]$$

$$\leq \Pr[\exists m' \in A : Q(m') > \frac{1}{4} Q(m) \text{ and } Q_{i-1}^{\text{err}} m' < \frac{1}{8} Q(m)]$$

$$+ \Pr[Q_{i-1}^{\text{err}} m + \frac{1}{t_i - 1} > 2 \cdot Q(m) \mid t_i \geq \frac{t}{2K}] + \Pr[t_i \leq \frac{t}{2K}]$$

$$+ \Pr[\exists m' \in A : Q(m') \leq \frac{1}{4} Q(m) \text{ and } Q_i^{\text{err}} m' > \frac{1}{2} Q(m)]$$

$$+ \Pr[Q_i^{\text{err}} m < \frac{1}{2} Q(m)] .$$

We bound each term individually. First, since $\{Q_{i-1}^{\text{err}} m \mid t_i\}$ (i.e., $Q_{i-1}^{\text{err}} m$ conditioned on a fixed $t_i$), $t_i$ and $Q_i^{\text{err}} m$ are binomial random variables with probabilities $Q(m)$, $\frac{1}{K}$ and $\frac{Q(m)}{K}$ respectively, the terms (II), (III) and (V) are easily upper bounded using the Chernoff bound. In particular,

$$\Pr[Q_{i-1}^{\text{err}} m + \frac{1}{t_i - 1} > 2 \cdot Q(m) \mid t_i \geq \frac{t}{2K}] \leq \exp \left(-\Omega \left(\frac{t \cdot Q(m)}{K}\right)\right)$$

$$\Pr[t_i \leq \frac{t}{2K}] \leq \exp \left(-\Omega \left(\frac{t \cdot Q(m)}{K}\right)\right)$$

$$\Pr[Q_i^{\text{err}} m < \frac{1}{2} Q(m)] \leq \exp \left(-\Omega \left(\frac{t \cdot Q(m)}{K}\right)\right) .$$
Term (I) is also upper bounded by Chernoff bound and a union bound over $m'$, since the number of $m'$ satisfying $Q(m') > \frac{1}{4}Q(m)$ is at most $4/Q(m)$. Thus,

$$\Pr \left[ \exists m' \in A : Q(m') > \frac{1}{4}Q(m) \text{ and } Q_i^t(m') < \frac{1}{8}Q(m) \right] \leq \frac{4}{Q(m)} \cdot \exp(-\Omega(t \cdot Q(m))).$$

To bound term (IV), we can assume without loss of generality that there is at most one $m' \in A$, such that, $Q(m') \leq \frac{1}{8}Q(m)$. This is because, if there were to exist $m'_1, m'_2 \in A$, such that, $Q(m'_1), Q(m'_2) \leq \frac{1}{8}Q(m)$, then we can identify $m'_1$ and $m'_2$ as the same message $m'_0$. Note that we can do this because we will still have that $\Pr[Q(m'_0) \leq \frac{1}{4}Q(m)]$ and

$$\Pr \left[ Q_i^t(m'_1) > \frac{1}{2}Q(m) \text{ or } Q_i^t(m'_2) > \frac{1}{2}Q(m) \right] \leq \Pr \left[ Q_i^t(m'_0) > \frac{1}{2}Q(m) \right]$$

Thus, to bound term (IV), we can again use a Chernoff bound and a union bound over $m'$, since the number of $m'$ such that $Q(m') > \frac{1}{8}Q(m)$ is at most $8/Q(m)$. Thus, we get that

$$\Pr \left[ \exists m' \in A : Q(m') \leq \frac{1}{4}Q(m) \text{ and } Q_i^t(m') > \frac{1}{2}Q(m) \right] \leq \frac{1}{Q(m)} \cdot \exp \left( -\Omega \left( t \cdot \frac{Q(m)}{K} \right) \right)$$

Thus, overall in any individual round, we have that,

$$p_{err}^t \leq \frac{1}{Q(m)} \cdot \exp \left( -\Omega \left( t \cdot \frac{Q(m)}{K} \right) \right).$$

This concludes the proof of Lemma 6.3.3.

6.5 Conclusion & Future Directions

We believe that the model introduced in this chapter is quite pertinent to the analyses of collective distributed phenomena where distributed entities are trying to come together to form joint actions. We believe the process and notation permit a much richer study, especially when one starts to allow correlations between the messages generated and the sender-receiver pairs. The ability to study the encoding and decoding functions — are they really functions, are they inverses of each other, how do they evolve? — are all intriguing questions that can now be subject to analyses. While our results do not address all these aspects, we do hope it will be the subject of future work.

In terms of the constructions and results, one interesting aspect of our compression protocol is that it mimics some of the curious features shown in human language. For every message $m$, player $i$ and round $t$, the encoding function describes a specific word which is player $i$’s encoding of $m$, i.e., it gives a (encoding) dictionary. The same player also possesses at the same round a decoding dictionary which we may view as saying, for every message $m$, which words this player would decode to $m$. Unlike in the basic schemes described, in our scheme the encoding dictionary is not identical to the decoding dictionary. While the
encoding dictionary is a function mapping messages to words, the decoding dictionary is not: It is more conservative and lists many words for any given message. This phenomenon is definitely visible in human languages and our work in this chapter suggests a plausible reason for the occurrence of this phenomenon.

We now mention some important questions that arise from this chapter. On the conceptual side, it would be very interesting to further use the formalism and ideas developed in theoretical computer science over the last decades in order to capture the phenomena exhibited by human languages. In particular, it would be interesting to extend our model to take into account other objectives along with compression. It would also be very interesting to consider the case where $Q$ and the set of interacting players vary (slightly) with time, in the hope of modelling “cultural” changes that take place from one generation to another.

On the more technical side, we stuck in this work to the complete graph representing the interactions between various players. It would be worthwhile to investigate other graph structures that favor the creation of communities, and study the properties of the language(s) that evolve in this case. Moreover, while we have considered in this work generic distributions $Q$, it would be nice to explore the data-structural aspects in the case where $Q$ comes from a well-structured family of distributions (e.g., a Markov Chain). Finally, a concrete question is to determine whether the $O(\log(D(P\|Q)))$ additive term in Theorem 6.3.1 is actually needed, which seems to be related to some intriguing questions about the chromatic number of certain families of graphs (see [HS14b]).
Chapter 7

Correlated Sampling

7.1 Introduction & Related Work

In this chapter, we study correlated sampling, a very basic task, variants of which have been considered in the context of sketching algorithms [Bro97], approximation algorithms based on rounding linear programming relaxations [KT02, Cha02], the study of parallel repetition [Hol07, Rao11, BHH+08] and very recently cryptography [Riv16].

This problem involves two players, Alice and Bob, attempting to come to agreement non-interactively. Alice and Bob are given distributions $P$ and $Q$ respectively over the same universe $\Omega$. Without any interaction, Alice is required to output an element $i \sim P$ and Bob is required to output an element $j \sim Q$, where the players have access to shared randomness. The goal is to minimize the disagreement probability $\Pr[i \neq j]$ in terms of the total variation distance $d_{TV}(P,Q)$ (where the probability is over the shared randomness). More formally, we define correlated sampling strategies as follows.

**Definition 7.1.1 (Correlated Sampling Strategies).** Given a universe\(^1\) $\Omega$ and a randomness space $R$, a pair of functions $(f, g)$, where $f: \Delta_\Omega \times R \rightarrow \Omega$ and $g: \Delta_\Omega \times R \rightarrow \Omega$, is said to be a correlated sampling strategy with error $\epsilon: [0,1] \rightarrow [0,1]$, if for any distribution $P, Q \in \Delta_\Omega$, such that $d_{TV}(P,Q) = \delta$, it holds that

- **[Correctness]** $\{f(P,r)\}_{r \sim R} = P$ and $\{g(Q,r)\}_{r \sim R} = Q$,

- **[Error guarantee]** $\Pr_{r \sim R} [f(P,r) \neq g(Q,r)] \leq \epsilon(\delta)$.

Here, $\Delta_\Omega$ is the set of all probability distributions on $\Omega$. Also, we abuse notations slightly to let $R$ denote a suitable distribution on the set $\mathcal{R}$. Moreover, we will always assume that $\mathcal{R}$ is sufficiently large, and we will often not mention $\mathcal{R}$ explicitly when talking about correlated sampling strategies. It is also allowed to have a sequence of strategies with increasing size of $\mathcal{R}$ in which case we want the above constraints to be satisfied in the limit as $|\mathcal{R}| \rightarrow \infty$.

\(^1\)we will primarily consider only finite universes.
A priori it is unclear whether such a protocol can even exist, since the error $\epsilon$ is not allowed to depend on the universe $\Omega$. Somewhat surprisingly, there exists a simple protocol whose disagreement probability can be bounded by roughly twice the total variation distance (and in particular does not degrade with the size of the universe). Variants of this protocol have been rediscovered multiple times in the literature yielding the following theorem.

**Theorem 7.1.2** (Holenstein [Hol07]. See also Broder [Bro97], Kleinberg-Tardos [KT02]). For any universe $\Omega$, there exists a correlated sampling strategy with error $\epsilon : [0, 1] \rightarrow [0, 1]$ such that

$$\forall \delta \in [0, 1], \quad \epsilon(\delta) \leq \frac{2 \cdot \delta}{1 + \delta}. \quad (7.1)$$

Strictly speaking, the work of Broder [Bro97] does not consider the general correlated sampling problem. Rather, it gives a strategy (the “MinHash strategy”) which happens to solve the correlated sampling problem under the condition that $P$ and $Q$ are flat distributions, i.e., they are uniform over some subset of the domain. The above bound applies to the case where these sets have the same size. The technique can also be generalized to other distributions to get the bound above, and this gives a protocol similar to that of Holenstein, though if $P$ and $Q$ are uniform over different sized subsets, the above bound is weaker than that obtained from a direct application of Broder’s algorithm! Holenstein [Hol07] appears to be the first to formulate the problem for general distributions and give a solution with the bound claimed above.

For the sake of completeness, we give a description of Broder’s strategy as well as Holenstein’s strategy in Section 7.3. We point out that variants of the protocol in Theorem 7.1.2 (sometimes referred to as “consistent sampling” protocols) had been used in several applied works [M+94, GP06, MMT10] (some of them before Holenstein’s paper).

Given Theorem 7.1.2, a natural and basic question is whether the bound on the disagreement probability can be improved. Indeed, this question was very recently raised by Rivest [Riv16] in the context of symmetric encryption, and this was one of the motivations behind this chapter. We give a surprisingly simple proof that the bound in Theorem 7.1.2 is actually tight (for a coarse parametrization of the problem)!

**Theorem 7.1.3** (Main Result). For every $\delta \in (0, 1)$ and $\gamma > 0$, there exists a family of pairs of distributions $(P, Q)$ satisfying $d_{TV}(P, Q) \leq \delta$ such that any correlated sampling strategy for this family has error at least $\frac{2 \delta}{1 + \delta} - \gamma$.

Our proof of Theorem 7.1.3 is surprisingly simple and is based on studying the following constrained agreement problem that we introduce and which is tightly related to correlated sampling. Alice is given a subset $A \subseteq [n]$ and Bob is given a subset $B \subseteq [n]$, where the pair $(A, B)$ is sampled from some distribution $D$. Alice is required to output an element $i \in A$ and Bob is required to output an element $j \in B$, such that the disagreement probability $\Pr_{(A,B) \sim D}[i \neq j]$ is minimized.

**Definition 7.1.4** (Constrained Agreement Strategies). Given a universe $\Omega = [n]$ and a distribution $D$ over $2^\Omega \times 2^\Omega$ (i.e., pairs of subsets of $\Omega$), a pair of functions $(f, g)$, where
\( f : 2^\Omega \to \Omega \) and \( g : 2^\Omega \to \Omega \) is said to be a constrained agreement strategy with error \( \text{err}_D(f, g) = \epsilon \in [0, 1] \), if it holds that

- [Correctness] \( \forall A \subseteq \Omega \), \( f(A) \in A \) and \( \forall B \subseteq \Omega \), \( g(B) \in B \),

- [Error guarantee] \( \Pr_{(A, B) \sim D} [f(A) \neq g(B)] \leq \epsilon \).

We point out that since the constrained agreement problem is defined with respect to an input distribution \( D \) on pairs of sets, we can require, without loss of generality, that the strategies \( (f, g) \) be deterministic in Definition 7.1.4 (this follows from Yao’s minimax principle). We arrive at the constrained agreement problem as follows: First we consider the flat distribution case of Definition 7.1.1 and relax the restrictions of \( \{f(P, r)\}_{r \sim \mathcal{R}} = P \) and \( \{g(Q, r)\}_{r \sim \mathcal{R}} = Q \) although we still require that \( f(P, r) \in \text{supp}(P) \) and \( g(Q, r) \in \text{supp}(Q) \) for any \( r \in \mathcal{R} \). This makes it a constraint satisfaction problem and we consider a distributional version of the same.

In order to prove Theorem 7.1.3, we show that in fact the correlated sampling strategy (a suitable de-randomization thereof) as in Theorem 7.1.2 is optimal for the constrained agreement problem whenever \( D \) is the distribution \( D_p \) where every coordinate \( i \in [n] \) is independently included in each of \( A \) and \( B \) with probability \( p \).

**Lemma 7.1.5.** For every \( p \in [0, 1] \) and for the distribution \( D_p \) on \( 2^{[n]} \times 2^{[n]} \), any constrained agreement strategy \((f, g)\) makes error \( \text{err}_{D_p}(f, g) \geq \frac{2(1-p)}{2-p} \).

**Organization.** In Section 7.1.1, we discuss some special cases of the correlated sampling problem. In Section 7.1.2, we give some open problems regarding these special cases. In Section 7.2, we prove Lemma 7.1.5 and use it to prove Theorem 7.1.3. In Section 7.3, we describe the correlated sampling protocols of Broder and Holenstein, thereby proving Theorem 7.1.2.

### 7.1.1 Special Cases

Let \( A, B \subseteq [n] \) be such that \( |A| = |B| \), and consider the problem of correlated sampling with the uniform distributions \( P = \mathcal{U}(A) \) and \( Q = \mathcal{U}(B) \). Then, the total variation distance between \( P \) and \( Q \) is given by

\[
d_{TV}(P, Q) = \frac{1}{2} \cdot \|P - Q\|_1 = \frac{1}{2} \cdot \|\mathcal{U}(A) - \mathcal{U}(B)\|_1 = 1 - \frac{|A \cap B|}{|A|}.
\]

Thus, the error probability of the correlated sampling strategy (in Theorem 7.1.2) is given by

\[
\frac{2 \cdot d_{TV}(P, Q)}{1 + d_{TV}(P, Q)} = 1 - \frac{|A \cap B|}{|A \cup B|}.
\] (7.2)

Rather surprisingly, in the particular case where \( |A \cap B| = 1 \) and \( A \cup B = [n] \), Rivest [Riv16] recently gave a protocol with smaller error probability than the one guaranteed by the correlated sampling protocol of Theorem 7.1.2.
Theorem 7.1.6 ([Riv16]). In Definition 7.1.1, if \( \Omega = [n] \), and the distributions \( P \) and \( Q \) are such that there exist \( A, B \subseteq [n] \) such that \( |A| = |B| \), \( |A \cap B| = 1 \), \( A \cup B = [n] \), and \( P = U(A) \) and \( Q = U(B) \), then, there is a correlated sampling strategy with error probability at most \( 1 - 1/|A| \).

For completeness, we describe this strategy in Section 7.3.1. Note that for this setting of parameters, we have that

\[
1 - \frac{1}{|A|} < 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{1}{n},
\]

and hence Theorem 7.1.6 improves on the performance (eq. (7.2)) of the correlated sampling strategy of Theorem 7.1.2. This naturally leads to the question: Can one similarly improve on the well-known correlated sampling protocol for larger intersection sizes, for example, when \( |A \cap B| \) is a constant fraction of \( |A| \)? The proof of our main result (Theorem 7.1.3) answers this question negatively. Namely, it implies that the strategy in Theorem 7.1.2 is tight when \( |A \cap B| = \epsilon \cdot |A| \) with \( \epsilon \in (0, 1) \) being an absolute constant.

Note that in the extreme case where \( \epsilon \) is very close to \( 0 \), Rivest’s protocol (Theorem 7.1.6) implies that Theorem 7.1.2 is not tight. What about the other extreme where \( \epsilon \) is very close to \( 1 \)? We show that in this case Theorem 7.1.2 is in fact tight.

Theorem 7.1.7. Let \( A, B \subseteq [n] \) be such that \( A \cup B = [n] \) and \( |A| = |B| = |A \cap B| + 1 \), and let \( P = U(A) \) and \( Q = U(B) \). Then, the error probability of any correlated sampling strategy is at least \( 1 - |A \cap B|/|A \cup B| \).

We prove Theorem 7.1.7 in Section 7.2.1.

7.1.2 Open Questions and Future Work

Our work started with a conjecture due to Rivest [Riv16] which informally asserts that Broder’s MinHash strategy is optimal except in the case considered in Theorem 7.1.6. More formally,

Conjecture 7.1.8 (Rivest [Riv16]). For every collection of positive integers \( n, a, b, \ell \) with \( \ell \geq 2 \) and \( n \geq a + b - \ell \), and for every pair of probabilistic strategies \((f, g)\) that satisfy correctness as in Definition 7.1.1, there exist \( A, B \subseteq [n] \) with \( |A| = a \), \( |B| = b \) and \( |A \cap B| = \ell \) such that

\[
Pr[f(A) \neq g(B)] \geq 1 - \frac{\ell}{a + b - \ell} = 1 - \frac{|A \cap B|}{|A \cup B|}.
\]

This chapter does not resolve this conjecture for the general setting of \( n, a, b \) and \( \ell \). It suggests this answer may be asymptotically right as \( n \to \infty \) when \( a = \alpha n \), \( b = \beta n \) and \( \ell = \alpha \beta n \), but does not exactly resolve this setting (our set sizes are only approximately \( \alpha n \)).

Even in the setting where the set sizes are allowed to vary slightly, our knowledge is somewhat incomplete. Lemma 7.1.5 shows optimality of the MinHash strategy when \((A, B) \sim D_p\).
In this case, $A$ and $B$ are independent and $p$-biased each, so $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \approx p^2 \cdot n$. We point out that a simple reduction to Lemma 7.1.5 also implies the optimality of the well-known protocol in the case where $A$ and $B$ are "positively-correlated". Specifically, consider the following distribution $\mathcal{D}_{p, \delta}$ on pairs $(A, B)$ of subsets of $[n]$, where we first sample $S \subseteq [n]$ which independently includes each element of $[n]$ with probability $p/(1 - \delta)$, and then independently includes every $i \in S$ in each of $A$ and $B$ with probability $1 - \delta$. In this case, $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \approx (1 - \delta) \cdot p \cdot n$. Even if we reveal $S$ to both Alice and Bob, Lemma 7.1.5 implies a lower bound of $2 \cdot \delta/(1 + \delta)$ on the error probability, which is achieved by the MinHash strategy. It is not clear how to use a similar reduction to show optimality in the case where $A$ and $B$ are "negatively-correlated", i.e., when $|A| \approx p \cdot n$, $|B| \approx p \cdot n$ and $|A \cap B| \ll p^2 \cdot n$.

Finally, the fact that Holenstein’s strategy for correlated sampling can be improved upon in the case where $P$ and $Q$ are uniform distributions on different-sized subsets of the universe clearly shows that strategy as in Theorem 7.1.2 is not “always optimal”. To study questions like this, one could restrict the class of pairs $(P, Q)$ and then give an optimal strategy for every $P$ and every $Q$. It would be interesting to study what would be the right measure that captures the minimal error probability given the adjacency relationship $(P, Q)$.

### 7.2 Lower Bounds on Correlated Sampling

We start by proving lower bounds on the error probability in the constrained agreement problem.

**Proof of Lemma 7.1.5.** Let $p \in [0, 1]$ and consider the distribution $\mathcal{D}_p$ on pairs $(A, B)$ of subsets $A, B \subseteq [n]$ where for each $i \in [n]$, we independently include $i$ in each of $A$ and $B$ with probability $p$. Let $f$ be Alice’s strategy which satisfies the property that $f(A) \in A$ for every $A \subseteq [n]$. Similarly, let $g$ be Bob’s strategy which satisfies the property that $g(B) \in B$ for every $B \subseteq [n]$.

We will construct functions $f^*$ and $g^*$ such that

$$\text{err}_{\mathcal{D}_p}(f, g) \geq \text{err}_{\mathcal{D}_p}(f^*, g) \geq \text{err}_{\mathcal{D}_p}(f^*, g^*) \geq \frac{2(1 - p)}{2 - p}.$$  

For every $i \in [n]$, we define $\beta_i \triangleq \text{Pr}_B[g(B) = i]$. Since under the distribution $\mathcal{D}_p$, the subsets $A$ and $B$ are independent, we have that when Bob’s strategy is fixed to $g$, the strategy of Alice that results in the largest agreement probability is given by

$$\forall A \subseteq [n], \quad f^*(A) = \arg \max_{i \in A} \beta_i.$$  

Thus, for a permutation $\sigma$ of $[n]$ such that $\beta_{\sigma^{-1}(1)} \geq \beta_{\sigma^{-1}(2)} \geq \cdots \geq \beta_{\sigma^{-1}(n)}$, we have that

$$\forall A \subseteq [n], \quad f^*(A) = \arg \min_{i \in A} \sigma(i).$$  

193
Now, for every \( i \in [n] \), we define \( \alpha_i \triangleq \Pr_A[f^*(A) = i] \). When Alice’s strategy is fixed to \( f^* \), the strategy of Bob that results in the largest agreement probability is given by

\[
\forall B \subseteq [n], \quad g^*(B) = \arg \max_{i \in B} \alpha_i.
\]

We now claim that \( \alpha_{\sigma^{-1}(1)} \geq \alpha_{\sigma^{-1}(2)} \geq \cdots \geq \alpha_{\sigma^{-1}(n)} \), and hence,

\[
\forall B \subseteq [n], \quad g^*(B) = \arg \min_{i \in B} \sigma(i).
\]

This follows easily because for each \( i \in [n] \), we have that,

\[
\alpha_i = \Pr_A \left( \arg \min_{i \in A} \sigma(\ell) = i \right) = (1 - p)^{i-1} \cdot p
\]

Thus, we conclude that

\[
\Pr_{(A,B) \sim \mathcal{D}_p}[f(A) = g(B)] \leq \Pr_{(A,B) \sim \mathcal{D}_p}[f^*(A) = g(B)] \leq \sum_{i=1}^{n} \Pr_{(A,B) \sim \mathcal{D}_p}[f^*(A) = g^*(B) = i]
\]

\[
= \sum_{i=1}^{n} \Pr_A[f^*(A) = i] \cdot \Pr_B[g^*(B) = i]
\]

\[
= \sum_{i=1}^{n} (1 - p)^{2(i-1)} \cdot p^2 \leq \frac{p}{2 - p},
\]

where the second equality uses the fact that under \( \mathcal{D}_p \), the subsets \( A \) and \( B \) are independent. Thus, we obtain that,

\[
\text{err}_{\mathcal{D}_p}(f,g) \geq 1 - \frac{p}{2 - p} = \frac{2(1 - p)}{2 - p}.
\]

We are now ready to prove our main result which is a lower bound on the error probability in correlated sampling.

**Proof of Theorem 7.1.3.** Let \( \delta \in (0,1) \) and \( \gamma > 0 \). Assume for the sake of contradiction that there is a correlated sampling strategy \( (f^*, g^*) \) that, when run on distributions at total
variation distance up to \( \delta \), has error probability at most \( \frac{2 \delta}{1 + \delta} - \gamma \). Fix \( \delta' \in (0, 1) \) such that

\[
\frac{2 \cdot \delta}{1 + \delta} - \gamma < \frac{2 \cdot \delta'}{1 + \delta'} < \frac{2 \cdot \delta}{1 + \delta}.
\]  (7.3)

Note that Equation (7.3) implies that \( \delta' < \delta \). Consider the distribution \( D_p \) over pairs \( (A, B) \) of subsets \( A, B \subseteq [n] \) where each \( i \in [n] \) is independently included in each of \( A \) and \( B \) with probability \( p \triangleq 1 - \delta' \). We then have that \( \mathbb{E}(|A|) = \mathbb{E}|B| = p \cdot n \), and \( \mathbb{E}|A \cap B| = p^2 \cdot n \). Moreover, by the Chernoff bound, we have that

\[
\Pr_A[||A| - p \cdot n| > p \cdot n^{0.99}] \leq e^{-p \cdot n^{0.98}/2},
\]

\[
\Pr_B[||B| - p \cdot n| > p \cdot n^{0.99}] \leq e^{-p \cdot n^{0.98}/2},
\]

and

\[
\Pr_{A,B}[|A \cap B| - p^2 \cdot n| > p^2 \cdot n^{0.99}] \leq e^{-p^2 \cdot n^{0.98}/2}.
\]

Hence, by the union bound and since \( p \leq 1 \), we get that with probability at least \( 1 - 3 \cdot e^{-p^2 \cdot n^{0.98}/2} \), we have that \( ||A| - p \cdot n| \leq p \cdot n^{0.99} \), \( ||B| - p \cdot n| \leq p \cdot n^{0.99} \) and \( ||A \cap B| - p^2 \cdot n| \leq p^2 \cdot n^{0.99} \). Consider now the distributions \( P = \mathcal{U}(A) \) (on Alice’s side) and \( Q = \mathcal{U}(B) \) (on Bob’s side). Then, with probability at least \( 1 - 3 \cdot e^{-p^2 \cdot n^{0.98}/2} \), it holds that

\[
d_{TV}(P, Q) = 1 - \frac{|A \cap B|}{\max\{|A|, |B|\}} \leq 1 - p + o_n(1) = \delta' + o_n(1) < \delta \text{ for sufficiently large } n.
\]

Note that Yao’s minimax principle implies that any correlated sampling strategy for \((P, Q)\) pairs with \( P = \mathcal{U}(A) \) and \( Q = \mathcal{U}(B) \) yields a constrained agreement strategy \((f, g)\) for the corresponding pairs \((A, B)\) of subsets. Hence, Lemma 7.1.5 implies that

\[
\forall f, g, \Pr_{(A, B) \sim D}[f(A) \neq g(B)] \geq \frac{2(1 - p)}{2 - p} = \frac{2 \cdot \delta'}{1 + \delta'},
\]  (7.4)

where \((f, g)\) is any correlated sampling strategy. On the other hand, the property of the assumed strategy \((f^*, g^*)\) implies that

\[
\exists f, g, \Pr_{(A, B) \sim D^*}[f(A) \neq g(B)] \leq \frac{2 \cdot \delta}{1 + \delta} - \gamma + o_n(1).
\]  (7.5)

Putting Equations (7.4) and (7.5) together contradicts Equation (7.3) for sufficiently large \( n \).

\[ \square \]
7.2.1 Lower Bound in a Special Case

In this section, we describe the lower bound in Theorem 7.1.7, which is incomparable to Theorem 7.1.3.

Proof of Theorem 7.1.7. Let $A, B \subseteq [n]$ be such that $|A| = |B| = |A \cap B| + 1$ and let $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$. Assume for the sake of contradiction that there is a correlated sampling strategy with disagreement probability $< 1 - |A \cap B|/|A \cup B| = 2/n$. Let $\mathcal{D}$ be the uniform distribution over pairs $(A, B)$ of subsets of $[n]$ satisfying $A \cup B = [n]$ and $|A| = |B| = |A \cap B| + 1$. Then, there is a deterministic strategy pair $(f, g)$ solving constrained agreement over $\mathcal{D}$ with error probability

$$\Pr_{(A, B) \sim \mathcal{D}}[f(A) \neq g(B)] < \frac{2}{n}. \quad (7.6)$$

Let

$$i \overset{\text{def}}{=} \arg \max_{\ell \in [n]} \left| \left\{ A \in \left[ \begin{array}{c} n \\ n-1 \end{array} \right] : f(A) = \ell \right\} \right|$$

be the element that is most frequently output by Alice’s strategy $f$, and denote its number of occurrences by

$$k \overset{\text{def}}{=} \left| \left\{ A \in \left[ \begin{array}{c} n \\ n-1 \end{array} \right] : f(A) = i \right\} \right|.$$

We consider three different cases depending on the value of $k$:

(i) If $k \leq n - 3$, then consider any $B \subseteq [n]$ with $|B| = n - 1$. For any value of $f(B) \in B$, the conditional error probability $\Pr[f(A) \neq g(B) \mid B]$ is at least $2/(n - 1)$. Averaging over all such $B$, we get a contradiction to Equation (7.6).

(ii) If $k = n - 2$, let $A_1 \neq A_2$ be the two subsets of $[n]$ with $|A_1| = |A_2| = n - 1$ such that $f(A_1) \neq i$ and $f(A_2) \neq i$. For any $B \subseteq [n]$ with $|B| = n - 1$ such that $B \neq A_1$ and $B \neq A_2$, the conditional error probability $\Pr[f(A) \neq g(B) \mid B]$ is at least $2/(n - 1)$. Note that there are $n - 2$ such $B$’s, and that either $A_1$ or $A_2$ is the set $[n] \setminus \{i\}$. If $B = [n] \setminus \{i\}$, then the conditional disagreement probability $\Pr[f(A) \neq g(B) \mid B]$ is at least $(n - 2)/(n - 1)$. Averaging over all $B$, we get that

$$\Pr_{(A, B) \sim \mathcal{D}}[f(A) \neq g(B)] \geq \left( \frac{1}{n - 1} \right) \cdot \left( \frac{n - 2}{n} \right) + \left( \frac{1}{n - 1} \right) \cdot \left( \frac{1}{n} \right) \geq \frac{2}{n},$$

where the last inequality holds for any $n \geq 2$. This contradicts Equation (7.6).

(iii) If $k = n - 1$, then the only subset $A_1$ of $[n]$ with $|A_1| = n - 1$ and such that $f(A_1) \neq i$ is $A_1 = [n] \setminus \{i\}$. For any $B \neq A_1$, the conditional error probability $\Pr[f(A) \neq g(B) \mid B]$ is
Protocol 6 MinHash strategy [Bro97]

Alice’s input: $A \subseteq [n]$
Bob’s input: $B \subseteq [n]$
Shared randomness: a random permutation $\pi : [n] \to [n]$

Strategy:

- $f(A, \pi) = \pi(i_A)$, where $i_A$ is the smallest index such that $\pi(i_A) \in A$.
- $g(B, \pi) = \pi(i_B)$, where $i_B$ is the smallest index such that $\pi(i_B) \in B$.

at least $1/(n-1)$. On the other hand, if $B = A_1$, then the conditional error probability is equal to $1$. Averaging over all $B$, we get that

$$\Pr_{(A,B) \sim \mathcal{D}}[f(A) \neq g(B)] \geq \left(\frac{1}{n-1}\right) \cdot \left(\frac{n-1}{n}\right) + 1 \cdot \left(\frac{1}{n}\right) = \frac{2}{n},$$

which contradicts Equation (7.6).

\[\square\]

7.3 Correlated Sampling Strategies

In this section, we describe the correlated sampling strategy that proves Theorem 7.1.2. First, let’s consider the case of flat distributions where the distributions $P$ and $Q$ are promised to be of the special form that there exist $A, B \subseteq [n]$ such that $P = \mathcal{U}(A)$ and $Q = \mathcal{U}(B)$ over the universe $[n]$. In this case, it is easy to show that the protocol given in Protocol 6 achieves an error probability of $1 - |A \cap B|/(A \cup B)$. Since $\pi$ is a random permutation, it is clear that $f(A, \pi)$ is uniformly distributed over $A$ and $g(B, \pi)$ is uniformly distributed over $B$. Let $i_0$ be the smallest index such that $\pi(i_0) \in A \cup B$. The probability that $\pi(i_0) \in A \cap B$ is exactly $|A \cap B|/(A \cup B)$, and this happens precisely when $f(A, \pi) = g(B, \pi)$. Hence, we get the claimed error probability.

The strategy desired in Theorem 7.1.2 can now be obtained by a reduction to the case of flat distributions, and subsequently using the MinHash strategy.

Proof of Theorem 7.1.2. Given a universe $\Omega$, define a new universe $\Omega' = \Omega \times \Gamma$, where $\Gamma = \{0, \gamma, 2\gamma, \ldots, 1\}$ for a sufficiently small value of $\gamma > 0$. Thus, $|\Omega'| = \frac{1}{\gamma} \cdot |\Omega|$. Suppose we are given distributions $P$ and $Q$ such that $d_{TV}(P, Q) = \delta$. Define $A = \{(\omega, p) \in \Omega \times \Gamma : p < P(\omega)\}$ and $B = \{(\omega, q) \in \Omega \times \Gamma : q < Q(\omega)\}$.
Holenstein’s strategy can now be simply described as follows: Alice and Bob use the MinHash strategy on inputs $A$ and $B$ over the universe $\Omega'$, to obtain elements $(\omega_A, p_A)$ and $(\omega_B, p_B)$ respectively, and they simply output $\omega_A$ and $\omega_B$ respectively. This strategy is summarized in Protocol 7.

It can easily seen that $|A| = \sum_{\omega \in \Omega} \left( \frac{P(\omega)}{\gamma} - 1 \right)$ and hence,

$$\sum_{\omega \in \Omega} \left( \frac{P(\omega)}{\gamma} - 1 \right) \leq |A| \leq \sum_{\omega \in \Omega} \frac{P(\omega)}{\gamma}.$$ 

Therefore,

$$\frac{1}{\gamma} - |\Omega| \leq |A|, |B| \leq \frac{1}{\gamma}.$$ 

Similarly, $|A \cap B| = \sum_{\omega \in \Omega} \min \left\{ \left\lfloor \frac{P(\omega)}{\gamma} \right\rfloor, \left\lfloor \frac{Q(\omega)}{\gamma} \right\rfloor \right\}$ and $|A \cup B| = \sum_{\omega \in \Omega} \max \left\{ \left\lfloor \frac{P(\omega)}{\gamma} \right\rfloor, \left\lfloor \frac{Q(\omega)}{\gamma} \right\rfloor \right\}$

and hence,

$$\frac{1 - \delta}{\gamma} - |\Omega| \leq |A \cap B| \leq \frac{1 - \delta}{\gamma},$$

$$\frac{1 + \delta}{\gamma} - |\Omega| \leq |A \cup B| \leq \frac{1 + \delta}{\gamma}.$$ 

The probability that Alice outputs $\omega_A$ is $\frac{P(\omega_A)}{|A|}$, which is bounded as

$$P(\omega_A) - \gamma \leq \frac{P(\omega_A)}{|A|} \leq \frac{P(\omega_A)}{1 - \gamma \cdot |\Omega|}.$$ 

Thus, it follows that as $\gamma \to 0$, Alice’s output is distributed according to $P$, and similarly Bob’s output is distributed according to $Q$. Moreover, we have that,

$$\Pr[\omega_A \neq \omega_B] = 1 - \frac{|A \cap B|}{|A \cup B|} \leq 1 - \frac{1 - \delta - \gamma \cdot |\Omega|}{1 + \delta} = \frac{2\delta + \gamma \cdot |\Omega|}{1 + \delta} \to \frac{2\delta}{1 + \delta}.$$ 

This gives us the desired error probability. 

\[ \square \]

### 7.3.1 Strategy in a Special Case

In this section, we describe the correlated sampling strategy of [Riv16] that proves Theorem 7.1.6. To do so, we will need the well-known Hall’s Theorem.

**Theorem 7.3.1 (Hall; [vLW01]).** Consider a bipartite graph $G$ on vertex sets $L$ and $R$. Then, there is a matching that entirely covers $L$ if and only if for every subset $S \subseteq L$, we have that $|S| \leq |N_G(S)|$, where $N_G(S)$ denotes the set of all neighbors of elements of $S$ in the graph $G$. 

198
Protocol 7 Holenstein’s strategy [Hol07]

Alice’s input: \( P \in \Delta_\Omega \)
Bob’s input: \( Q \in \Delta_\Omega \)

Pre-processing: Let \( \Omega' = \Omega \times \Gamma \), where \( \Gamma = \{0, \gamma, 2\gamma, \cdots, 1\} \) (for suitable \( \gamma > 0 \))

Shared randomness: \( r \sim \mathcal{R} \) as required by the MinHash strategy on \( \Omega' \)

Strategy:

- Let \( A = \{(\omega, p) \in \Omega \times \Gamma : p < P(\omega)\} \) and \( B = \{(\omega, q) \in \Omega \times \Gamma : q < Q(\omega)\} \).
- Alice and Bob use MinHash strategy (Algorithm 6) with inputs \( A, B \) on universe \( \Omega' \) to obtain \( (\omega_A, p_A) \) and \( (\omega_B, p_B) \) respectively.
- Alice outputs \( \omega_A \).
- Bob outputs \( \omega_B \).

Proof of Theorem 7.1.6. Alice and Bob have subsets \( A, B \subseteq [n] \) respectively such that \( |A| = |B| = k \), \( |A \cap B| = 1 \) and \( A \cup B = [n] \). This forces \( n = 2k - 1 \). Consider the bipartite graph \( G \) on vertices \( [n] \times [n] \), with an edge between vertices \( A \) and \( B \) if \( |A \cap B| = 1 \). It is easy to see that \( G \) is \( k \)-regular. Iteratively using Hall’s theorem (Theorem 7.3.1), we get that the edges of \( G \) can be written as a disjoint union of \( k \) matchings. Let’s denote these as \( M_1, M_2, \cdots, M_k \).

The strategy of Alice and Bob is as follows: Use the shared randomness to sample a random index \( r \in [k] \) and consider the matching \( M_r \). If \( (A, B') \) is the edge present in \( M_r \), then Alice outputs the unique element in \( A \cap B' \). Similarly, if \( (A', B) \) is the edge present in \( M_r \), then Bob outputs the unique element in \( A' \cap B \). This protocol is summarized in Protocol 8.

It is easy to see that both Alice and Bob are outputting uniformly random elements in \( A \) and \( B \) respectively. Moreover, the probability that they output the same element, is exactly \( 1/k \), which is the probability of choosing the unique matching \( M_r \) which contains the edge \( (A, B) \) (i.e. enforcing \( A = A' \) and \( B = B' \)).
Protocol 8 Rivest’s strategy [Riv16]

Alice’s input: $A \subseteq [n]$
Bob’s input: $B \subseteq [n]$
Promise: $|A| = |B| = k$, $|A \cap B| = 1$ and $A \cup B = [n]$

Pre-processing: Let $G$ be the bipartite graph on vertices $\binom{[n]}{k} \times \binom{[n]}{k}$, with an edge between vertices $A$ and $B$ if $|A \cap B| = 1$. Decompose the edges of $G$ into $k$ disjoint matchings $M_1, \ldots, M_k$.

Shared randomness: Index $r \in [k]$

Strategy:

- Let $(A, B')$ and $(A', B)$ be edges present in $M_r$.
- Alice outputs the unique element in $A \cap B'$.
- Bob outputs the unique element in $A' \cap B$. 


Chapter 8

Conclusion

In this thesis, we studied various aspects of uncertainty in communication, including the contextual components related to the goal of the communication, the randomness shared by the parties and the prior distribution of the inputs. Our work leaves open several fascinating questions, both technical and conceptual.

Technical Questions A very interesting question raised by our work on common randomness and secret-key generation is whether one can obtain time-efficient decoders for our explicit sample-efficient schemes (see Section 2.11 for more details and related open questions).

The main questions raised by our work on the non-interactive simulation of joint distributions are to obtain computational hardness results, and to extend our techniques for proving decidability to other tensor-power problems such as computing the zero-error Shannon capacity of a graph (see Section 3.9 for more details and related open questions).

Our work on communication with functional uncertainty leads to several interesting technical questions on the power of interaction, private-randomness and imperfectly shared randomness in uncertain communication. In particular, can we obtain an efficient protocol in the case where the functions are promised to have small two-way communication without uncertainty? See Section 5.11 for more details and related open questions.

The main technical open question related to our work on uncertain distributed compression is to reduce the communication cost down to the entropy of the unknown distribution without increasing the number of iterations. As mentioned in Section 6.5, doing so seems to require answering some open questions related to the chromatic number of certain families of graphs.

Conceptual Questions In the communication with functional uncertainty setting, it would be interesting to study finer measures of distance between the players’ contexts than the Hamming distance that we considered.

Another question is to further study the computational complexity aspects in both communication with functional uncertainty and uncertain distributed compression.
An ideal model for communication should only assume a constant amount of perfectly shared context between the sender and receiver, such as the knowledge of an encoding/decoding algorithm, one universal Turing machine, etc.. Solutions to most interesting communication problems seem to assume a shared information which grows with the length of the inputs. Some of our work shows that in many of these scenarios some assumptions about the shared context can be relaxed to an imperfect sharing, but these results are often brittle and break when two or more contextual elements are simultaneously assumed to be imperfectly shared. For instance, our work raises the question of whether imperfectly shared randomness would be sufficient to overcome functional uncertainty. We show that this is indeed the case for product distributions, but the loss for non-product distributions might be much larger (see Chapter 5 and Section 5.11 for more details). Such results highlight the delicate nature of the role of shared context in communication. They beg for a more systematic study of communication which at the very least should be able to mimic the aims, objectives and phenomena encountered in human communication.

It would also be interesting to investigate whether the work done in natural language processing (e.g., on conversational AI) can be connected to our models of communication with uncertainty, and whether some of our solutions can shed a new light on some of the questions studied in that domain.

Within theoretical computer science, communication complexity has been extensively studied and several strong lower bound results have been proved and used, with great success, to obtain negative results in several other areas (such as streaming algorithms, data structures, etc.). It would be interesting if some of these tools can be further used to study communication as an end in itself, which is of paramount importance both in the human case and in the case of conversing intelligent systems.
Bibliography


