ÉCOLE POLYTECHNIQUE DE BRUXELLES

UNIVERSITÉ LIBRE DE BRUXELLES

Quantum Weak Coin Flipping

where weakness is a virtue

Thesis presented by Atul Singh ARORA

with a view to obtaining the PhD Degree in Engineering Sciences ("Docteur en Sciences de l'ingénieur et technologie")

Academic year 2019-2020

Supervisor: Professor Jérémie ROLAND

Centre for Quantum Information and Communication (QuIC)

Thesis jury:

Nicolas CERF (Université libre de Bruxelles, Chair)
Stefano PIRONIO (Université libre de Bruxelles, Secretary)
Peter HØYER (University of Calgary)
André CHAILLOUX (Institut national de recherche en sciences et technologies du numérique)
Professor Samuel FIORINI (Université libre de Bruxelles)

fnrs LA LIBERTÉ DE CHERCHER

Thesis submitted with a view to obtaining the PhD Degree in Engineering Sciences

Quantum Weak Coin Flipping

where weakness is a virtue

Atul Singh ARORA

October 2016–August 2020

Submitted to the Jury:	April 20, 2020
Private Defence:	June 25, 2020
Revision:	August 21, 2020
Public Defence:	August 25, 2020

Université libre de Bruxelles
Ecole polytechnique de Bruxelles
Centre for Quantum Information and Communication

Supervisor:
Professor Jérémie ROLAND

Thesis jury:
Professor Nicolas CERF (Université libre de Bruxelles, Chair)
Professor Stefano PIRONIO (Université libre de Bruxelles, Secretary)
Professor Peter HØYER (University of Calgary)
Professor André CHAILLOUX (Institut national de recherche en sciences et technologies du numérique)
Professor Samuel FIORINI (Université libre de Bruxelles)

Academic year:
2019–2020

Funded by:
Bourse FRIA (Fonds pour la formuation à la Recherche dans l'Industrie et dans l'Agriculture),
FRS-FNRS Fonds de la Recherche Scientifique.

With characteristic modesty Gauss declared that

"If others would but reflect on mathematical truths as deeply and as continuously as I have, they would make my discoveries."

Possibly. Gauss' explanation recalls Newton's.

Asked how he had made discoveries in astronomy surpassing those of all his predecessors, Newton replied,

"By always thinking about them."

This may have been plain to Newton; it is not to ordinary mortals.

Bell, E.T.. Men of Mathematics (Touchstone Book) (p. 254). Touchstone. Kindle Edition.

Contents

List of Figures

Acknowledgments

This got out of control very quickly. I leave it for the amusement of the reader; admittedly, not the best form of entertainment.

List of people who directly contributed to the results.

- Jérémie ROLAND
- Stephan WEIS
- Chrysoula VLACHOU

People who fell victim to the Lord of Coins.

- Nicolas CERF
- Ognyan ORESHKOV
- Mathieu BRANDEHO
- Tom VAN HIMBEECK
- Stefano PIRONIO
- Kishor BHARTI

Those who fell victim to my existence and were not already listed.

- Levon CHAKMAKCHAN
- Alexis HIBLER
- Matthieu ARNHEM
- Leonardo Goncalves NOVO
- Shantanav CHAKRABORTY
- Zacharie Van HERSTRAETEN
- Guy PAUL
- Michael JABBOUR
- Anaelle HERTZ
- Samara HUSSAIN
- Siddhartha DAS
- Uttam SINGH
- Evgueni KARPOV
- Luc VANBEVER
- Célia GRIFFET
- Timothée HOFFREUMON
- Jef PAUWELS
- Pascale LATHOUWERS

Strange encounters, fun friendships.

- Riccardo LONGO

- Quentin DELHAYE
- Erica BERGHMAN
- Marc CRUELLAS BORDES
- Loïc BLANC

Ones who also featured in my MS thesis's acknowledgment.

- Manu JAYADHARAN
- Srijit MUKHERJEE
- Vivek SAGAR
- Yosman BAPATDHAR
- Saumya GUPTA
- Evelyn ABRAHAM
- Shwetha SRINIVASAN

Finally, the select few who are listed separately to infuse a dramatic touch—people I am perhaps the most indebted to.

- Kishor BHARTI
- Ritu ROY CHOWDHURY
- Vivek SAGAR
- Yaiza GALLARDO
- Family—my sisters, my father, my mothers (biological and otherwise)
- Belgian Family—Jordi TIÓ ROTLLAN, Maria Isabel NOGUERAS VILA, Gérard TIÓ NOGUERAS, Carla TIÓ NOGUERAS; Giulia GAGGIOTTI, Zohreh CHAHARDOLI, Daniel CASAS

Abstract

We investigate weak coin flipping, a fundamental cryptographic primitive where two distrustful parties need to remotely establish a shared random bit. A cheating party can try to bias the output bit towards a preferred value. For weak coin flipping the parties have known opposite preferred values. By a weak coin flipping protocol with bias ϵ we mean that neither player can force the outcome towards their preferred value with probability more than $1/2 + \epsilon$. While it is known that classically, $\epsilon = 1/2$ (the worst possible), Mochon showed in 2007 that quantumly, weak coin flipping can be performed with arbitrarily small bias (near perfect). His non-constructive proof used the so-called point game formalism—a series of equivalent reductions which were introduced by Kitaev to study coin-flipping. He constructed point games with bias $\epsilon_M(k) = 1/(4k + 2)$ to prove the existence. The best known explicit protocol, however, had bias approaching $\epsilon_M(1) = 1/6$ (also due to Mochon, 2005).

In the present work, we try to make the non-constructive part of the proof constructive, to wit, we make three main contributions towards the conversion of point-games into explicit protocols. First, we propose a framework—TIPG-to-Explicit-protocol Framework (TEF)—which simplifies the task of constructing explicit protocols. We use this framework to construct a protocol with bias $\epsilon_M(2) = 1/10$. We then give the exact formulae for the unitaries corresponding to the point-games due to Mochon, allowing us to describe (almost) perfect coin flipping protocols analytically, i.e. with bias $\epsilon_M(k)$ for arbitrarily large k. Finally, we introduce an algorithm we call the Elliptic Monotone Align (EMA) algorithm. This algorithm, together with TEF, lets us convert any point-game into an explicit protocol numerically. We conclude by giving another analytic construction of unitaries for Mochon's games using the ellipsoid picture introduced for the EMA algorithm.

Resumé

Nous étudions le *weak coin flipping*, une primitive cryptographique fondamentale où deux parties méfiantes doivent établir à distance un bit aléatoire partagé. Un tricheur peut essayer de biaiser le bit de sortie vers une valeur préférée. Pour le *weak coin flipping*, les parties ont des valeurs préférées opposées. Par un protocole de *weak coin flipping* avec biais ϵ, nous entendons qu'aucun des deux joueurs ne peut forcer le résultat vers sa valeur préférée avec une probabilité supérieure à $1/2 + \epsilon$. Alors que l'on sait que classiquement, $\epsilon = 1/2$ (le pire possible), Mochon a montré en 2007 qu'un *weak coin flipping* quantique peut être effectué avec un biais arbitrairement faible (presque parfait). Sa preuve non constructive a utilisé le formalisme dit du jeu de points (*point games*)—une série de réductions équivalentes qui ont été introduites par Kitaev pour étudier le coin flipping. Il a construit des jeux de points avec un biais $\epsilon_M(k) = 1/(4k + 2)$ pour en prouver l'existence. Le protocole explicite le plus connu, cependant, avait un biais approchant $\epsilon_M(1) = 1/6$ (également dû à Mochon, 2005).

Dans le présent travail, nous essayons de rendre la partie non constructive de la preuve constructive, c'est-à-dire que nous apportons trois contributions principales à la conversion des jeux de points en protocoles explicites. Premièrement, nous proposons un cadre—*TIPG-to-Explicit-protocol Framework (TEF)*—qui simplifie la tâche de construction de protocoles explicites. Nous utilisons ce cadre pour construire un protocole avec un biais $\epsilon_M(2) = 1/10$. Nous donnons ensuite les formules exactes des unitaires correspondant aux jeux de points dus à Mochon, ce qui nous permet de décrire analytiquement des protocoles de *coin flipping* (presque) parfaits, c'est-à-dire avec un biais $\epsilon_M(k)$ pour un k arbitrairement grand. Enfin, nous introduisons un algorithme que nous appelons le *Elliptic Monotone Align (EMA) Algorithm*. Cet algorithme, associé à *TEF*, nous permet de convertir numériquement tout jeu de points en un protocole explicite. Nous concluons en donnant une autre construction analytique des unitaires pour les jeux de Mochon en utilisant l'image ellipsoïdale introduite pour l'algorithme *EMA*.

List of Publications

- "Quantum weak coin flipping". With: Jérémie Roland and Stephan Weis. In: *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing - STOC 2019*. ACM Press. DOI: 10.1145/3313276.3316306. arXiv:1811.02984.
  - In this thesis, this result comprises Chapter 4 and Chapter 6.
- "Analytic quantum weak coin flipping protocols with arbitrarily small bias". With: Jérémie Roland and Chrysoula Vlachou. In: *Submitted*. arXiv:1911.13283.
  - In this thesis, this result comprises Chapter 5. The first version of this preprint comprised Chapter 7.
- "Uniqueness of all fundamental noncontextuality inequalities". With: Kishor Bharti, Leong Chuan Kwek, and Jérémie Roland. In: *Phys. Rev. Research 2, 033010 (July 2020)*. DOI: 10.1103/PhysRevResearch.2.033010. arXiv:1811.05294.
  - Unrelated to this thesis.
- "Revisiting the admissibility of non-contextual hidden variable models in quantum mechanics". With: Kishor Bharti and Arvind. In: *Physics Letters A Volume 383, Issue 9 (February 2019)*. Elsevier. DOI:10.1016/j.physleta.2018.11.049. arXiv:1607.03498.
  - Unrelated to this thesis. Continuation of my master's thesis.

1

Chapter

Introduction

The problem studied in this thesis is rather easy to state. Suppose there are two parties, conventionally called Alice and Bob, who are placed in physically remote locations and can communicate with each other using a communication channel. They wish to exchange messages over this channel in order to agree on a random bit. To be more concrete, further suppose they want to use the random bit to decide who must read this thesis.[1] This is easy to do—Alice flips a coin and messages Bob the outcome. However, this requires Bob to trust Alice. Can Bob modify the scheme to be sure that Alice did not cheat? More generally, can one construct a protocol (which involves an exchange of messages over the communication channel) to decide on a random bit while ensuring that an honest party, i.e. one that follows the protocol, can not be deceived? It turns out that if one communicates over a classical communication channel—such as the internet or the telephone—then a cheating party can always force their desired outcome on the honest party unless we make further assumptions, such as computational hardness. If, on the other hand, Alice and Bob are allowed to use a quantum communication channel, then protocols solving this problem, up to vanishing errors, have been shown to *exist*. We say exist because the proof of this statement, which was a breakthrough result from 2007, has one non-constructive step. This means, rather interestingly, that we know that there is a solution to the problem but we don't know the solution itself. In this thesis, we build upon the previous pioneering works to solve quantum weak coin flipping, as this problem is referred to in the literature, and betrayed by the title of the thesis. We define it more precisely in Section 1.2.

The problem also occupies an interesting place in the overall landscape of cryptography (see Section 1.1). It was shown in 1994 that a public key cryptosystem, called RSA, which is widely used even today can be broken using a quantum computer [35]. A decade earlier, a method for key distribution using quantum channels was proposed [6] whose security, in principle, relied only on the validity of the laws of physics. These developments, together with Grover's search algorithm [17], resulted in the now fiercely growing field of investigation, collectively termed Quantum Information. It was thought that quantum mechanics could also revolutionise secure two-party computations, another branch of cryptography (see Subsection 1.1.2). Success here, was marred by a cascade of impossibility results. In a breakthrough result of (classical) cryptography, it was shown that a primitive called Oblivious Transfer (OT) is universal for secure two-party computations. An analogy could be that the NAND gate is universal for all functions. It is known that no protocol for OT can exist that uses only classical messages and offers perfect security without relying on further assumptions, such as computational hardness. OT deprived quantum mechanics of being the panacea for cryptography—it was shown that even if the communication is quantum, OT can not be implemented with perfect security. Perhaps one can be less ambitious and target Bit Commitment (BC), a cryptographic primitive weaker than OT. Researchers were again disappointed. BC is also impossible (in the same sense) even with a quantum communication channel. Coin flipping, an even weaker primitive suffered the same fate. Weak coin flipping (WCF), however, was poised for fame. It is the strongest known primitive (in the secure two-party setting) which admits no secure classical protocol but can be implemented over a quantum channel with near perfect security.

[1] and make the reasonable assumption that nobody wishes to read it

While an explicit WCF protocol has remained evasive, ingenious connections have been discovered. In particular, it was shown that if one can magically perform WCF, then using this ability and only classical communication, one can implement optimal strong flipping, i.e. (asymptotically) the most secure strong coin flipping protocol permissible by the laws of quantum mechanics. Using WCF as a subroutine, it was shown that one can also implement optimal BC and a variant of OT.

The most significant advance in the study of WCF was the invention of the so-called point games, attributed to Alexei Kitaev. He showed that there are three equivalent formalisms which can be used to describe WCF protocols—explicit protocols (pairs of dual SDPs), Time Dependent Point Games (TDPGs), and Time Independent Point Games (TIPGs). The fact that WCF with almost perfect security is quantumly possible, was established by Mochon using the TIPG formalism. The main difficulty in going backwards, i.e. to an explicit protocol, was that no constructive method was known for obtaining an explicit protocol from a TDPG. In this thesis, we use perturbative methods to describe a protocol with bias $1/10$ while the best known protocol had a bias of $1/6$ (also due to Mochon, 2005).[2] We then introduce a more systematic method for converting Mochon's point games into explicit unitaries (which in turn are readily converted into explicit protocols) by identifying an appropriate decomposition of the main "move" used in the point game. This technique was tailored for Mochon's point game and *a priori* not expected to work in general. To address this, we introduce a numerical algorithm that allows one to provably perform the non-constructive step in the conversion of a TDPG into an explicit protocol. This, in effect, lets one numerically construct WCF protocols corresponding to any TIPG, including Mochon's which approach zero bias. Finally, we give another analytic solution to Mochon's point games which is inspired by the techniques used in the numerical solution.

Readers familiar with the main results in the two-party secure computation setting (using quantum communication), may prefer skipping Section 1.1 below. In the subsequent section, Section 1.2, we define the coin flipping problem more precisely and prove that it doesn't admit a classical protocol. The proof is concise and bears the seeds of its quantum generalisation. We also introduce the various formalisms in some detail. This allows us to, in Section 1.3, informally describe our technical contributions, as outlined above.

§ 1.1 Cryptography muddled with quantum results

This section combines parts of the lecture notes by Rafael Pass and Abhi Shelat, titled "A course in Cryptography", and those by Stephanie Wehner and Thomas Vidick titled "EdX: Quantum Cryptography".

1.1.1 Overview of the field

Historically, the field of *cryptography*, the art of secure communication, may be understood as following a *crytpo-cycle*. (1) Alice, invents an encryption scheme. (2) Alice claims/proves security against known attacks. (3) The encryption scheme gets employed (often in critical situations). (4) The scheme is eventually broken by improved attacks. (5) Alice and their friends restart.

A major part of cryptography was really cryptanalysis, breaking known encryption techniques. The modern field of cryptography has evolved into one with the following aims: (1) Provide mathematical definitions of security, (2) Provide precise mathematical assumptions (e.g. "factoring is hard" where *hard* is formally defined), (3) Provide proofs of security, i.e. demonstrate that under some precise assumptions, the scheme cannot be broken.

[2] Strictly speaking, these are families of protocols whose bias approaches the said value asymptotically. In the informal discussions, we often suppress this subtlety to improve readability.

Cryptography goes beyond secure communication—exchanging (possibly sensitive) information over a public channel (of communication). Suppose Alice and Bob each have a (private) list and wish to find the intersection of the two lists without revealing the remaining contents of their lists. A real world example could be that two (or more) large financial banks wish to determine their "common risk exposure" without revealing anything else about their investments. Clearly, a trusted third party could be used to mediate such a task but would either bank be willing to trust the third party with their sensitive and valuable information? This task is a special case of what is known as a *secure two-party computation*. Details thereof comprise the following subsection.

To see where our problem of interest sits in a wider context, we briefly list some of the main pillars of modern cryptography. *Computational Hardness and One-way functions*—To circumvent Shannon's lower bound, i.e. the key must be as long as the message for perfect security, one must restrict ones attention to computationally-bounded adversaries. Most results rely on one-way functions assuming resource-bounds (in particular time bounds). A one-way function is easy to compute but hard to invert (easy/hard in the sense of efficiency). It is at the heart of modern cryptographic protocols. *Indistinguishability*—Given a computationally-bounded player, can two distributions be distinguished? It is central to defining security; it also pivotal for formal definitions of pseudo-random generators, commitment schemes, zero-knowledge protocols etc. *Knowledge*—Protocol execution shouldn't leak "knowledge". There are knowledge-based definitions of security. *Authentication*—Digital signatures and message authentication codes serve as digital analogues of traditional written signatures. *Computing on Secret Inputs*—Two mutually distrustful players wish to perform arbitrary computation on their (possibly secret) information. Includes secret-sharing protocols and secure two-party computation protocols. The latter we discuss shortly. Secret-sharing protocols allow n parties to receive "shares" of a secret such that any 'small' subset of shares leaks almost no information but enough of them together reveal the secret. *Composability*—A cryptographic scheme, which is secure in isolation, when run parallelly might become completely insecure. Composability tackles these issues.

1.1.2 Secure Two-Party Computation

Here, we develop the notion of secure two-party computation and then delve deeper into two primitives—Oblivious Transfer and Bit Commitment—which we mentioned in the introduction. Our objective is to introduce reasonably representative definitions for these and motivate the proofs of the basic results which relate them. This discussion is not pivotal to the understanding of the main results of the thesis and therefore some readers may prefer skipping directly to Section 1.2.

One notion of secure two-party computations may be formalised using what is known as *secure function evaluation (SFE)*. In these contexts, the definitions of security can be slightly involved and are not directly relevant to our work here. Thus, we satisfy ourselves with the following informal statement.

Definition 1 (Secure function evaluation (SFE))**.** Secure function evaluation involves two parties, Alice and Bob. Alice and Bob hold (possibly secret) inputs $x \in \mathscr{X}$ and $y \in \mathscr{Y}$, respectively. They interact over a communication channel and output $a \in \mathscr{A}$ and $b \in \mathscr{B}$ respectively. Let $f_A : \mathscr{X} \times \mathscr{Y} \to \mathscr{A}$, $f_B : \mathscr{X} \times \mathscr{Y} \to \mathscr{B}$ respectively be the functions they wish to evaluate. To any protocol evaluating these functions we ascribe the following properties, granted the associated condition holds:

- Correctness: If both Alice and Bob follow the protocol, i.e. they are both *honest*, then $a = f_A(x, y)$ and $b = f_B(x, y)$
- Security against cheating Bob: If Alice is honest, then Bob cannot learn more about her input than he can infer from $f_B(x.y)$
- Security against cheating Alice: If Bob is honest, then Alice cannot learn more about his input than she can infer from $f_A(x, y)$.

While the statement is quite straightforward, formalising the notion of "Bob cannot learn more about her input" it turns out is not trivial. Ignoring the subtlety, we consider a simple example: Alice and Bob have to decide on taking a questionable action, like watching a comedy film which is considered to be in poor taste. Suppose 0 denotes "not interested" and 1 denotes "interested". Denote Alice's preference by x and Bob's by y. If both are interested (i.e. $x = y = 1$), then the action is taken. However, if Alice is not interested (i.e. $x = 0$), Bob doesn't want her to know whether or not he was interested and *vice versa*. Formally, $f_A(x, y) = f_B(x, y)$, because they both must agree (nobody likes to be stood-up), which in turn must equal $x \wedge y$, the AND function.

In this section, we intuitively see an intriguing result, followed by various impossibility theorems. These motivate the study of coin-flipping and, to some extend, establish the fundamental nature of the problem as we already alluded to. We defer a more complete discussion to the end. To be able to state the result, we state an equally intriguing problem.

Definition 2 (1-out-of-2 Oblivious Transfer (1-2 OT))**.** Oblivious Transfer is an SFE problem. Let $\mathscr{X} = \{0,1\}^l \times \{0,1\}^l$ and $\mathscr{Y} = \{0,1\}$. Alice receives as input two l-bit strings, $x = (s_0, s_1) \in \mathscr{X}$ and Bob receives a single bit $y \in \mathscr{Y}$. They wish to evaluate $f_A(x, y) = \perp$ and $f_B(x, y) = s_y$ respectively, where $\mathscr{A} = \{\perp\}$ and $\mathscr{B} = \{0,1\}^l$.

Let us take a moment to understand this. Alice holds two inputs. Bob wants to learn one of them. A 1-2 OT protocol should allow Bob to learn one of Alice's inputs (and prevent any knowledge of the other) while simultaneously ensuring that Alice does not learn which of her two inputs was learnt by Bob, i.e. Alice remains *oblivious* to which input she *transferred* to Bob. This seemingly impossible-to-solve problem, is indeed, impossible quantumly (and so classically as well), unless one makes further assumptions (usually these assume bounds on computational power (such as time or space); restrictions on the communication channel; relaxed security guarantees). However, this is not just one SFE problem which can not be solved—all SFE problems are impossible (in the same sense). This is because 1-2 OT is known to be universal for all SFE problems (even multi-player variants [16]), much like the NAND gate is universal for computational problems (we already saw this analogy in the introduction).

Oblivious Transfer is Universal for Secure Function Evaluation

We now sketch the intuition behind the proof of universality of 1-2 OT. For simplicity, suppose $f_A = f_B = f$ (the argument trivially extends). Alice constructs a garbled circuit corresponding to f (an encrypted description of a circuit; details are introduced as needed), and knows for each wire, including the input wires, the relation between the encryption key and the wire[3] value 0 or 1. She sends the description of the garbled circuit to Bob, the "evaluating player", together with the encryption keys corresponding to her input x. Since she only sends the keys, the circuit will evaluate in accordance with the input x, but Bob has no way of learning x. Bob however, can not even input his own inputs, y, into the encrypted circuit. Worse still, he can not ask Alice for the corresponding keys because then Alice learns y (compromising the security). Enter: 1-2 OT. Bob and Alice perform a 1-2 OT which solves exactly this problem—Alice sends the key corresponding to Bob's input, without learning the input. Now Bob can evaluate the circuit, obtain $f(x, y)$ and send the result to Alice. If $f_A \neq f_B$ then Alice can construct a garbled circuit corresponding to f_B and proceed as above. Subsequently they switch roles, Bob constructs a garbled circuit corresponding to f_A and they proceed analogously.

Surprisingly, 1-2 OT was first provided (partly of course) by Wiesner [38] (in 1973), using a quantum communication channel. Only later, Rabin [32] (in 1981) constructed a computational variant and coined the term oblivious transfer.

[3] the encryption key corresponding to value 0 will allow the circuit to be evaluated with the wire set to 0, for instance

Relation between Oblivious Transfer and Bit Commitment

Bit commitment is yet another cryptographic problem. Suppose Alice wishes to *commit* to a bit b but she doesn't want Bob to learn b. Later, Alice should be able to send some additional information which *reveals* the bit to Bob. The protocol must ensure that Alice can not change her commitment, i.e. send some additional information which reveals a different bit. A helpful (almost cartoon) implementation is through a safe and a key. Suppose Alice writes her bit on a piece of paper, puts it in the safe and gives the safe the Bob. This is the commit phase. When she wishes to reveal, she sends Bob the key.[4] In this section, we construct a scheme for implementing 1-2 OT over a quantum channel, assuming bit-commitment (BC) can be implemented.[5] For completeness, we give a formal definition and a short remark but they may be skipped.

Definition 3 (Bit Commitment (BC))**.** Suppose Alice wants to commit her input $b \in \{0,1\}$ to Bob (who has no input). At the end, Alice outputs $\perp$ and Bob outputs $b' \in \{0,1,\perp\}$. A protocol for bit commitment has two phases: the first is the *commit* phase and the second is the *reveal* phase. The protocol must satisfy the following properties.

1. (Correctness) If both Alice and Bob are honest, then at the end of the protocol, Bob outputs $b' = b$.
2. (Hiding) For any malicious Bob, the state of Bob at the end of the commit phase is independent of b.
3. (Binding) Denote Alice's cheating strategy for the commit phase by A, and thereafter by $A_{0/1}$ depending on the outcome she wishes to reveal. Let p_b be the probability that Bob outputs $b' = b$ after being cheated by Alice (who uses strategies A and A_b). The binding property is that $p_0 + p_1 \leq 1$ for all possible strategies, A, A_0 and A_1.

Remark 4. The binding property tries to capture this: Once Alice has committed to a b (using strategy A), the she shouldn't be able to come up with two different strategies (A_0 and A_1) such that she can convince Bob that $b = 0$ (by running A_0) or that $b = 1$ (by running A_1) with probability (strictly) greater than $1/2$.

We now give a representative construction of 1-2 OT using BC. Using the notation from the problem statement as in Definition 2, except that we denote Alice's input by (s_0, s_1) thereby freeing the symbol x. The protocol is as follows.

1. Alice chooses strings $x \in \{0,1\}^{2n}$ and $\theta \in \{0,1\}^{2n}$ uniformly at random. She prepares the states $|x_j\rangle_{\theta_j}$ for each $j \in \{1, 2 \ldots 2n\}$ where

$$|0\rangle_0 = |0\rangle\,, \qquad |1\rangle_0 = |1\rangle$$
$$|0\rangle_1 = |+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \qquad |1\rangle_1 = |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}.$$

   She sends these states to Bob.

[4] Obviously, exchanging safes is not the idea; besides their security relies on mechanical difficulties and are thus vulnerable.

[5] Given 1-2 OT, one can implement BC using classical communication. Suppose Alice's wishes to commit to her input, $b \in \{0,1\}$ (while Bob has no input).

Commit: Bob chooses two random strings, $s_0, s_1 \in \{0,1\}^l$. Bob sends s_b to Alice, while being oblivious to which he sent, using 1-2 OT. By the end of this, Alice has s_b while Bob has both s_0, s_1 but doesn't know anything about b. (Even though it looks like Alice hasn't given anything to Bob, she has implicitly communicated b to obtain the means of convincing him later of her commitment, that is s_b)

Reveal: Alice sends $b' = b$ and the string $s'_{b'} = s_b$ to Bob. Bob verifies that $s'_{b'} = s_{b'}$, if this is the case, he outputs b' (concludes Alice had committed herself to b'). If not, i.e. if $s'_{b'} \neq s_{b'}$ then he outputs $\perp$ (concludes Alice did not commit).

Security: OT guarantees that Bob doesn't learn b until Alice reveals it. Alice may still fool Bob but with at most a negligible probability 2^{-l}.

2. Bob measures each state in the basis $\tilde{\theta}_j$ for $\tilde{\theta} \in \{0,1\}^{2n}$ chosen uniformly at random. Let the outcomes be $\tilde{x}_j$. He commits to $\{\tilde{\theta}_1, \tilde{\theta}_2 \dots \tilde{\theta}_{2n}\}$ and $\{\tilde{x}_1, \tilde{x}_2 \dots \tilde{x}_n\}$ using BC with Alice.
3. Alice picks a subset $R \subset \{0,1\}^{2n}$ and asks Bob to reveal $\tilde{\theta}_r$ and $\tilde{x}_r$ for all $r \in R$. If for any r, $\tilde{\theta}_r = \theta_r$ and $x_r \neq \tilde{x}_r$, Alice aborts, claiming Bob cheated.
(This tries to force Bob into actually doing the measurements)
4. Alice reveals her choice of basis $\{\theta_1, \theta_2 \dots \theta_{2n}\}$ to Bob.
(As Alice is convinced Bob has indeed measured, she can send the choice of basis)
5. (Recall Bob wished to learn the yth string, s_y.) Bob defines $I_y := \{i : \theta_i = \tilde{\theta}_i\} \backslash R$, i.e. the set of indices for which their basis choice matched (excluding the ones Alice used for testing Bob's commits) and $I_{1-y} := \{1, 2 \dots 2n\} \backslash I_y \cup R$, i.e. the set of indices for which their basis choice did not match (for simplicity, we suppose that $|I_0| = |I_1| = l = n$). Bob sends (I_0, I_1) to Alice.
(Bob is creating two sets of indices, from one he can learn Alice's input, from the other he can not; he sends these to Alice)
6. Alice sends $t_0 = s_0 \oplus x_{I_0}$ and $t_1 = s_1 \oplus x_{I_1}$ to Bob.
(Alice does not know from which set Bob can learn her inputs but she knows that from one of them, he can)
7. Alice outputs $\perp$ while Bob outputs $t_y \oplus \tilde{x}_{I_y}$.

The protocol is correct because when both players are honest, $\tilde{x}_i = x_i$ for $i \in I_y$ and so $t_y \oplus \tilde{x}_{I_y} = s_y \oplus x_{I_y} \oplus x_{I_y} = s_y$, so indeed, Bob obtains s_y. Intuitively, a cheating Alice cannot learn y because (I_0, I_1) do not contain any information about y as Bob never reveals the $\tilde{\theta}_i$s for $i \notin R$. A cheating Bob could not have postponed the measurements as Alice forces him to commit by running a test. He could, however, choose to send Alice indices which let him learn both s_0 and s_1 partially; this subtlety can be handled using privacy amplification which we do not discuss here.

We hope to have made plausible, by this discussion, the result that quantumly, bit commitment implies oblivious transfer which in turn implies secure function evaluation [13, 39]. This is an interesting result by itself because it is known that classically, bit commitment is strictly weaker than oblivious transfer. However, the quantum advantage does not take us much further.

Impossibility of Ideal Bit Commitment (information theoretic)

Ideal bit commitment is impossible (in the information theoretic setting), even quantumly [25, 24]. We give an informal argument. Consider any bit commitment protocol which is perfectly hiding. This would entail that during the commit phase, Bob holds a state $\rho_B(0)$ corresponding to Alice having committed to zero and $\rho_B(1)$ corresponding to her having committed to one. Perfect hiding entails that $\rho_B(0) = \rho_B(1)$, else Bob could distinguish the states with some non-zero success probability. One can assume that Alice holds a purification of the state, i.e. $|\psi_0\rangle_{AB}$ or $|\psi_1\rangle_{AB}$ which are such that $\rho_B(0) = \mathrm{tr}_B(|\psi_0\rangle_{AB} \langle\psi_0|_{AB})$ and $\rho_B(1) = \mathrm{tr}_B(|\psi_1\rangle_{AB} \langle\psi_1|_{AB})$. If we can show that Alice can transform $|\psi_0\rangle_{AB}$ into $|\psi_1\rangle_{AB}$ using a local unitary, i.e. $U_A \otimes \mathbb{I}_B$, then we would have established the impossibility of bit commitment. This is a well known result, called the Uhlmann's theorem.

Theorem 5 (Uhlmann's Theorem). *Given a density matrix ρ_B and its purification $|\psi\rangle_{AB}$, another state $|\phi\rangle_{AB}$ is also a purification of ρ_B if and only if there exists a unitary U_A such that $|\phi\rangle_{AB} = U_A \otimes \mathbb{I} |\psi\rangle_{AB}$.*

This statement is not very hard to prove if one uses the Schmidt decomposition (which itself is a clever application of singular value decomposition to a pure bipartite state) to write

$$|\psi\rangle_{AB} = \sum_i \lambda_i |u_i\rangle_A |v_i\rangle_B$$

$$|\phi\rangle_{AB} = \sum_i \mu_i |w_i\rangle_A |z_i\rangle_B$$

where the coefficients λ_i and μ_i are non-negative (real) numbers, indexed in the decreasing order. Suppose, for simplicity, there is no degeneracy, i.e. the eigenvalues are all distinct. Note that since $\mathrm{tr}_B(|\psi\rangle\langle\psi|) = \mathrm{tr}_B(|\phi\rangle\langle\phi|)$, we must have $\lambda_i = \mu_i$ and $|v_i\rangle = |z_i\rangle$. The only freedom left then, is in the choice of $|u_i\rangle_A$ and $|w_i\rangle_A$, i.e. a unitary freedom in system A, i.e. $U_A \otimes \mathbb{I}_B$.

To summarise, we saw that perfect bit-commitment is impossible, even quantumly, which entails perfect oblivious transfer is impossible,[6] which in turn entails that perfect secure function evaluation is impossible.

It was later also shown (see, for instance, [10]) that quantum bit commitment must necessarily have a finite amount of imperfection, thereby eliminating nearly perfect bit commitment as well. Is there a weaker primitive in this distrustful setting which we can implement quantumly but is impossible classically? We address this in the next section.

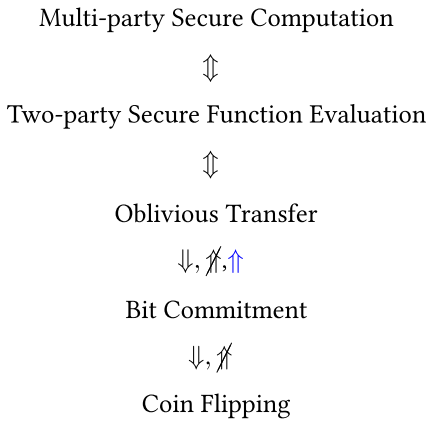


Figure 1.1: Relation between the various cryptographic primitives. Bit commitment and Oblivious Transfer are quantumly equivalent but not classically (hence the blue implication).

§ 1.2 Coin Flipping

In this section, we define a cryptographic primitive called coin flipping (both weak and strong), together with the notion of a bias for any protocol which implements it. Subsequently, we see why classically, coin flipping is impossible, following Kitaev's analysis. We also state a theorem (again, due to Kitaev) which entails that strong coin flipping is impossible, even quantumly. While the proof of the forrmer (which constitutes most of Subsection 1.2.2) is succinct and insightful, it is not a logical necssisity for the remainder of the thesis. It may thus be skipped.

1.2.1 Problem Statement

Coin flipping may be thought of as a two player game where the players communicate with each other and in the end, output $x, y \in \{0, 1\}$. Their communication may be over a quantum channel or over a classical channel. Denote by p_{xy} the probability that Alice outputted x and Bob outputted y. When both players are honest (i.e. follow the rules of the game) we have

$$p_{00} = p_{11} = \frac{1}{2}; \quad p_{01} = p_{10} = 0.$$

If Bob cheats (while Alice is honest), suppose Prob[Alice gets 0] $\in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right]$ and if Alice cheats (while Bob is honest), suppose Prob[Bob gets 0] $\in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right]$. Such a game, often called *strong coin flipping*, is impossible in the classical setting for any $\epsilon < \frac{1}{2}$; we show this using Kitaev's argument [23].

[6] One can prove the impossibility of OT directly as well, again, by using Uhlmann's theorem.

To this end, we define a weaker game by adding more conditions. Suppose Alice wants to bias the result towards 0 and Bob wants to bias the result towards 1 (and they are both aware of this). This game is usually termed *weak coin flipping*. Define $p_{x*} = \max_{\tilde{B}} \text{Prob}[\text{Alice gets } x]$ where the maximization over $\tilde{B}$ represents a maximization over all possible cheating strategies for Bob. Analogously, define $p_{*y} = \max_{\tilde{A}} \text{Prob}[\text{Bob gets } y]$. For weak coin flipping, we need only consider

$$p_{*0} \leq \frac{1}{2} + \epsilon; \quad p_{1*} \leq \frac{1}{2} + \epsilon$$

because Alice would not try to bias towards 1 and Bob would not try to bias towards 0. To contrast, suppose we considered a coin flipping game where the cheater wants the other player to output 1 (but Alice does not know Bob's intent and vice versa). In this case, we would impose $p_{*1} \leq \frac{1}{2} + \epsilon; \quad p_{1*} \leq \frac{1}{2} + \epsilon$.

1.2.2 Impossibility of Classical Coin Flipping and Quantum Strong Coin Flipping

We now state two important theorems due to Kitaev.

Theorem 6 (Kitaev's first CF theorem (informal))**.** *For both quantum and classical coin flipping games,* $p_{x*}p_{*y} \geq p_{xy}$.

We do not give the proof of the quantum part of this proof but sketch the proof of the classical part later. The following corollary is, however, easy to prove.

Corollary 7. *Strong coin flipping is impossible for any* $\epsilon \leq \frac{1}{\sqrt{2}} - \frac{1}{2}$ *(applies to both quantum and classical games).*

Proof. For strong coin flipping we must have $p_{*y} \leq \frac{1}{2} + \epsilon$ and $p_{x*} \leq \frac{1}{2} + \epsilon$ for all $x, y \in \{0, 1\}$. Using $x = y = 1$, in the theorem, we have $p_{1*}p_{*1} \geq p_{11} = \frac{1}{2}$ which implies $p_{1*} \geq \frac{1}{\sqrt{2}}$ or $p_{*1} \geq \frac{1}{\sqrt{2}}$. □

Suprisingly, it turns out that in the classical case, one can prove a stronger impossibility result.

Theorem 8 (Kitaev's second CF theorem (informal))**.** *For a classical coin flipping game,* $(1 - p_{x*})(1 - p_{*y}) \leq \sum_{x' \neq x, y' \neq y} p_{x'y'}$.

This statement rules out all non-trivial classical coin flipping games.

Corollary 9. *Classical weak coin flipping is impossible for any* $\epsilon < \frac{1}{2}$.

Proof. We simply substitute for $x = 1$ and $y = 0$ to obtain $(1 - p_{1*})(1 - p_{*0}) \leq p_{01} = 0$ which entails that either $p_{1*} = 1$ or $p_{*0} = 1$. □

Classically, there is no advantage to be had from weak coin flipping. Quantumly, we know that strong coin flipping must have a finite bias. What about quantum weak coin flipping? The remainder of this document is dedicated towards answering this question. We end this subsection by proving Theorem 8.

Proof sketch for (the classical part of) Theorem 6 and Theorem 8. Consider a classical game with N rounds (excluding the last step for the output), i.e. Alice sends a message/bit-string a_1 to Bob, then Bob sends a message b_2 to Alice and so on, until Bob sends a message b_{2N} to Alice (see Figure 1.2).

Define the state of the protocol at step $2k$ as $u(2k) = (a_1, b_2, a_3 \ldots b_{2k})$ and at $2k + 1$ as $u(2k + 1) = (a_1, b_2, a_3 \ldots a_{2k+1})$. The state at the beginning of the protocol is just the empty string $\emptyset$. The state at the end is $(u(2N), x, y)$. Define

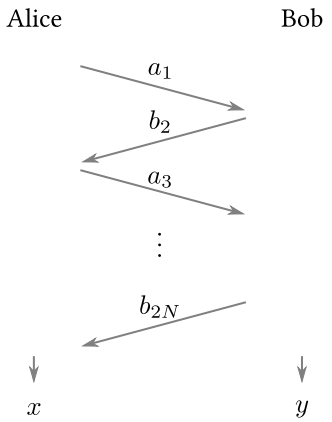


Figure 1.2: General classical coin flipping protocol with $2N$ rounds.

- $w(u)$ to be the probability of being in the state u in an honest game
- $Z^A(u)$ to be the maximum probability of Bob's success in cheating Alice (while Alice is honest)
- $Z^B(u)$ to be the maximum probability of Alice's success in cheating Bob (while Bob is honest)

The key idea, due to Kitaev which he later puts to extraordinary use in the quantum case for both weak and strong coin flipping (we only discuss the former), is to combine these three into a single object. He defines $F_j = \sum_{u \in \text{level}(j)} w(u) Z^A(u) Z^B(u)$ where by level j we mean after j steps. (We use the word level as one can imagine a tree of possible states.) The strategy is to consider the boundary conditions, i.e. F_0 and F_{2N+1} and to then establish a relation between F_i and F_{i+1}.

Boundary condition. We have $w(\emptyset) = 1$, $Z^A(\emptyset) = p_{x'*}$, $Z^B(\emptyset) = p_{*y'}$ which follow from the definitions of w, Z^A and Z^B. Thus,

$$F_0 = p_{x'*} p_{*y'}. \tag{1.1}$$

Suppose $u \in \text{level}(2N)$. A little thought also shows that $\sum_u w(u,x,y) = p(x,y)$. Further, $Z^A(u,x,y) = \delta_{x,x'}$ because the probability of Bob cheating and winning, at the last step, is either 1, i.e. he succeeded at forcing x', or 0, he failed. Analogously, $Z^B(u,x,y) = \delta_{y,y'}$. We therefore have

$$F_{2N+1} = \sum_{u \in \text{level}(u), x \in \{0,1\}, y \in \{0,1\}} w(u,x,y)(\delta_{xx'})(\delta_{yy'}) = p_{x'y'}. \tag{1.2}$$

Monotonicity. Suppose it is Alice's turn. Then, we can write

$$w(u,a) = p(a|u) w(u) \tag{1.3}$$

using Baye's rule for conditional probabilities. Recall that Z^A was the probability of success of a cheating Bob playing against an honest Alice. Since it is Alice's turn, she chooses a according to the conditional probability, $p(a|u)$. What is slightly harder to see is that

$$Z^A(u) = \sum_a p(a|u) Z^A(u,a). \tag{1.4}$$

We know that at level u, Alice would act according to the distribution $p(a|u)$. If we know that at the subsequent levels, the probability of Bob's success at cheating is $Z^A(u,a)$ we know that at the level before, that is at level u, his probability of success is the weighted average of his success in the subsequent level where the weight is governed by $p(a|u)$. What can we say about $Z^B(u)$? We have

$$Z^B(u) = \max_a Z^B(u,a) \tag{1.5}$$

because Z^B is the maximum success probability of a cheating Alice and we assumed it was Alice's turn. If she were in the state u, and she know the success probabilities associated with all the states (u, a), she would simply pick the one that maximized her probability of winning.

We can now prove the monotonicity result. We have

$$\begin{aligned}
\sum_a w(u,a)Z^A(u,a)Z^B(u,a) &= w(u)\sum_a w(a|u)Z^A(u,a)Z^B(u,a) && \text{using (1.3)}\\
&\leq w(u)Z^B(u)\sum_a w(a|u)Z^A(u,a) && \text{using (1.5)}\\
&= w(u)Z^A(u)Z^B(u) && \text{using (1.4).}
\end{aligned}$$

Summing over u, we have

$$F_{i+1} \leq F_i. \tag{1.6}$$

Combining Equation (1.1), Equation (1.2) and Equation (1.6), we conclude that $p_{x'y'} \leq p_{x'*}p_{y'*}$ which is Theorem 6, restricted to the classical case. This can also be derived more systematically using linear programming duality. The quantum version, based on semi-definite programming duality also goes through analogously. The proof of Theorem 8 is quite similar but relies on the fact that $0 \leq Z^{A/B} \leq 1$ which ceases to hold in the quantum case. To complete the classical proof, define

$$G_j := \sum_{u \in \text{level}(j)} w(u)(1 - Z^A(u))(1 - Z^B(u))$$

and note that we have $G_0 = (1 - p_{x'*})(1 - p_{*y'})$, $G_{2N+1} = \sum_{x\neq x', y\neq y'} p_{x,y}$ together with $G_{i+1} \geq G_i$. □

1.2.3 Quantum Weak Coin Flipping

We begin our discussion of quantum weak coin flipping by reviewing the state of the art, this time in a little more detail. This is followed by a description of the three main formalisms which are central to this subject. Our main results foundationally rely on these. Thus we spend some time to intuitively explain them, before proceeding further.

As we saw, classically coin flipping is impossible. While quantumly, for strong coin flipping (from Corollary 7) we concluded that the bias must at least be $\frac{1}{\sqrt{2}} - \frac{1}{2}$, the best known explicit protocol is due to Ambainis and has bias $\frac{1}{4}$ [4]. As for weak coin flipping (WCF), the current best known explicit protocol—the Dip Dip Boom protocol—is due to Mochon [27] and has bias $1/6$. In a breakthrough result, he even proved the existence of a quantum weak coin-flipping protocol with arbitrarily low bias $\epsilon > 0$, hence showing that near-perfect weak coin flipping is theoretically possible [28]. This fundamental result for quantum cryptography, unfortunately, was proved non-constructively, by elaborate successive reductions (80 pages) of the protocol to different versions of so-called point games, a formalism introduced by Kitaev [23] in order to study coin flipping. Consequently, the structure of the protocol whose existence is proved is lost. A systematic verification of this by independent researchers recently led to a simplified proof [3] (*only* 50 pages) but over a decade later, an explicit weak coin-flipping protocol is still unknown, despite various expert approaches ranging from the distillation of a protocol using the proof of existence to numerical search [30, 31]. Further, weak coin flipping provides, via black-box reductions, optimal protocols for strong coin flipping [12], bit commitment [10] and a variant of oblivious transfer [11] (fundamental cryptographic primitives). It is also used to implement other cryptographic tasks such as leader election [15] and dice rolling [1].

To understand the problem statement better, let us start with noting two features of weak coin flipping.[7] First, note that we can say, without loss of generality, that if the bit is zero it means Alice won and if the bit is one it means Bob won. Why is that? In weak coin flipping we know both players have known preferences. Alice wants zero and Bob wants one[8] since if they both wanted the same outcome bit, there would be no need to flip a coin. If a player gets what they want, we say they won. Second, we note that there are four situations which can arise in a weak coin flipping scenario of which three are of interest. Let us denote by HH the situation where both Alice and Bob are honest, i.e. follow the protocol. In this situation we want the protocol to be such that both Alice and Bob (a) win with equal probability and (b) are in agreement with each other. In the situation HC where Alice is honest and Bob is cheating, the protocol must protect Alice from a cheating Bob. In this situation, a cheating Bob tries to convince an honest Alice that he has won. His probability of succeeding by using his best cheating strategy is denoted by P_B^* where the star/asterisk refers to a cheating player and the subscript denotes the outcome he desires to enforce on the honest player. The CH situation where Bob is honest and Alice is cheating naturally points us to the corresponding definition of P_A^*. The situation CC where both players are cheating is not of interest to us as nothing can be said which depends on the protocol. This is because nobody is following the protocol.

A trivial example of a weak coin flipping protocol is where Alice flips a coin and reveals the outcome to Bob over the phone. A cheating Alice can simply lie and always win against an honest Bob which means $P_A^* = 1$. On the other hand, a cheating Bob can not do anything to convince Alice that he has won, unless it happens by random chance on the coin flip. This corresponds to $P_B^* = \frac{1}{2}$. The bias of the protocol is $\max[P_A^*, P_B^*] - \frac{1}{2}$ which for this naïve protocol amounts to $\frac{1}{2}$, the worst possible. Manifestly, constructing protocols where one player is protected is nearly trivial. Constructing protocols where neither player is able to cheat (against an honest player) is the real challenge.

Given a WCF protocol it is not a priori clear how the best success probability of a cheating player, denoted by $P_{A/B}^*$, should be computed as the strategy space can be dauntingly large. It turns out that all quantum WCF protocols can be defined using the exchange of a message register interleaved with the players applying the unitaries U_i locally (see Figure 1.3) until a final measurement, say Π_A denoting Alice won and Π_B denoting Bob won, is made in the end. Computing P_A^* in this case reduces to a semi-definite program (SDP) in ρ, the corresponding quantum state: maximise $P_A^* = \mathrm{tr}(\Pi_A \rho)$ given the constraint that the honest player follows the protocol. Similarly for computing P_B^* one can define another SDP. Using SDP duality one can turn this maximization problem over cheating strategies into a minimization problem over dual variables $Z_{A/B}$. Any dual feasible assignment then provides an upper bound on the cheating probabilities $P_{A/B}^*$. SDPs are usually easy to handle but in this case, there are two SDPs, and we must optimise both simultaneously (see Subsection 2.1). Note that here we assume the protocol is known and we are trying to find bounds on P_A^* and P_B^*. However, our goal is to find good protocols. So what we would like is a formalism which allows us to do both, construct protocols and find the associated P_A^* and P_B^*. Kitaev gave us such a formalism.

He converted this problem about matrices (Zs, ρs and Us) into a problem about points on a plane, which Mochon called Kitaev's Time Dependent Point Game (TDPG) formalism. In this formalism, one is concerned with a sequence of frames (also referred to as configurations)—the positive quadrant of the plane with some points and their probability weights—which must start with a fixed frame and end with a frame that has only one point. The fixed starting frame consists of two points at $[\![0, 1]\!]$ and $[\![1, 0]\!]$ with weight $1/2$. The end frame must be a single point, say at $[\![\beta, \alpha]\!]$, with weight 1. The objective of the protocol designer is to get this end point as close to the point $[\![\frac{1}{2}, \frac{1}{2}]\!]$ as possible by transitioning through intermediate frames (see Figure 1.4) by following certain rules (we describe these shortly). The magic of this formalism, roughly stated, is that if one abides by these rules then corresponding to every such valid sequence of frames, there exists a WCF protocol with $P_A^* = \alpha$, $P_B^* = \beta$ (see Subsection 3.1).

[7] While we already introduced some of the concepts, we choose to be slightly redundant to ensure the independent readability of this section.

[8] Alice wanting a zero and Bob wanting a one is just an uninteresting relabelling.

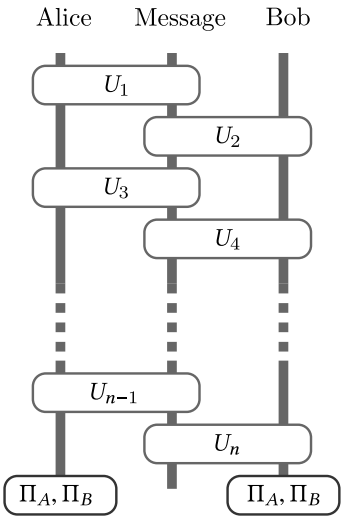


Figure 1.3: General structure of a Weak Coin Flipping protocol.

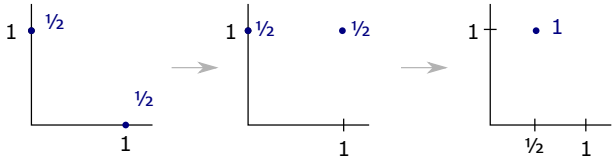


Figure 1.4: Point game corresponding to the weak coin flipping over the phone protocol.

We now describe these rules. Consider a given frame and focus on a set of points that fall on some vertical (or horizontal) line. Let the y coordinate (or x coordinate) of the ith point be given by z_{g_i} and the weight be given by p_{g_i}. Let z_{h_i} and p_{h_i} denote the corresponding quantity in the subsequent frame. Then, the following conditions[9] must hold

1. the probabilities are conserved, viz. $\sum_i p_{g_i} = \sum_i p_{h_i}$
2. for all $\lambda > 0$

$$\sum_i \frac{\lambda z_{g_i}}{\lambda + z_{g_i}} p_{g_i} \leq \sum_i \frac{\lambda z_{h_i}}{\lambda + z_{h_i}} p_{h_i}. \tag{1.7}$$

Note that from one frame to the next, one can either make a horizontal transition or a vertical transition. By combining these sequentially one can obtain the desired form of the final frame, i.e. a single point. The aforesaid rule and the points in the frames arise from the dual variables $Z_{A/B}$. Just as the state ρ evolves through the protocol, so do the dual variables $Z_{A/B}$. The points and their weights in the TDPG are exactly the eigenvalue pairs of $Z_{A/B}$ with the probability weight assigned to them by the honest state $|\psi\rangle$ at a given point in the protocol ($|\psi\rangle$ and ρ are closely related). The aforementioned rules are related to the dual constraints. Given an explicit WCF protocol and a feasible assignment for the dual variables witnessing a given bias, it is straightforward to construct the TDPG. However, going backwards, constructing the WCF dual from a TDPG is non-trivial and no general construction is known.

[9] To be precise, the condition $\sum_i p_{g_i} z_{g_i} = \sum_i p_{h_i} z_{h_i}$ must also be added but it only affects the limiting behaviour (of derived quantities such as strictly validity) which we can neglect in this discussion.

As this point game formalism is the cornerstone of the analysis, we simplify the rules further and then apply them to construct a simple example game. Later, we convert this example game into an explicit protocol using our . Before proceeding, we take a moment to formalise the aforementioned description into a *transition*. A transition is a finitely supported function from the interval, $[0, \infty)$ to itself. The transition from the given frame to the next is written as $\sum_i p_{g_i} [\![z_{g_i}]\!] \to \sum_i p_{h_i} [\![z_{h_i}]\!]$ where the function

$$[\![a]\!](x) = \delta_{a,x}, \tag{1.8}$$

i.e. it is zero when $x \neq a$ and one when $x = a$. The corresponding *function*[10] is written as $t = \sum_i p_{h_i} [\![z_{h_i}]\!] - \sum_i p_{g_i} [\![z_{g_i}]\!]$. If the transition/function satisfies the conditions (1) and (2), it is termed a *valid transition/function*.

If we restrict ourselves to transitions involving only one initial and one final point, i.e. $[\![z_g]\!] \to [\![z_h]\!]$, the second condition reduces to $z_g \leq z_h$ (we suppressed the subscript). This is called a *raise*. It means that we can always increase the coordinate of a single point. What about going from one initial point to many final points, i.e. $[\![z_g]\!] \to \sum_i p_{h_i} [\![z_{h_i}]\!]$ (note that the points before and after must lie along either a horizontal or a vertical line)? The second condition in this case becomes $1/z_g \geq \langle 1/z_h \rangle$, that is the harmonic mean[11] of the final points must be greater than or equal to that of the initial point (which in turn equals the coordinate of the initial point as there is only one initial point), where $\langle f(z_h) \rangle := \left(\sum_i f(z_{h_i}) p_{h_i}\right) / \left(\sum_j p_{h_j}\right)$. This is called a *split*. Finally, we can ask: What happens upon merging many points into a single point, i.e. $\sum_i p_{g_i} [\![z_{g_i}]\!] \to [\![z_h]\!]$? The second condition becomes $\langle z_g \rangle \leq z_h$, that is the final position must not be smaller than the average initial position (where $\langle f(z_g) \rangle$ is analogously defined). This is called a *merge*. While these three transitions/moves do not exhaust the set of moves, they are enough to construct games that almost achieve the bias $1/6$. Let us construct a simple game as an example. We start with the initial frame and raise the point $[1, 0]$ along the vertical to $[1, 1]$ (see Figure 1.4). We know this move is allowed as it is just a raise. Next we merge the points $[0, 1]$ with $[1, 1]$ using a horizontal merge. The x-coordinate of the resulting point can at best be $\frac{1}{2}.0 + \frac{1}{2}.1 = \frac{1}{2}$ where we used the fact that both points have weight $1/2$. Thus we end up with a single point at $[\frac{1}{2}, 1]$ with all the weight. Kitaev's formalism tells us that there must exist a protocol which yields $P_A^* = 1$ while $P_B^* = \frac{1}{2}$. This, however, is the phone protocol that we started our discussion with! It is a neat consistency check but it yields a trivial bias. This is because we did not use the split. If we use a split once, we can, by essentially matching the weights, already obtain a game with $P_A^* = P_B^* = \frac{1}{\sqrt{2}}$. Protocols corresponding to this bias were found by various researchers [37, 29, 22] long before this formalism was known. In fact, the bias of the said weak coin flipping protocol, $\epsilon = \frac{1}{\sqrt{2}} - \frac{1}{2}$, was exactly the lower bound for *strong* coin flipping. It was an exciting time (we imagine) as the technique used to bound strong coin flipping fails for weak coin flipping. The matter was not resolved for a while. This protocol held the record for being the best known weak coin flipping protocol until Mochon progressively showed that if we use multiple splits wisely at the beginning followed by a raise, one simply needs to use merges thereafter to obtain a game with bias almost $1/6$, which corresponds to his Dip Dip Boom protocol. The Dip Dip Boom protocol, is actually a family of protocols which in the limit of infinite rounds of communication yields bias $1/6$. Going lower, therefore, is not a straight forward extension and moves which can not be decomposed into the three basic ones: splits, merges and raises, are needed.

§ 1.3 Contributions (informal)

We present our contributions almost chronologically with one exception. We state the exact solution at the earliest permitted by logical dependencies, to allow the reader to reach a description of explicit

[10] details about overlaps and thus cancellations are suppressed for simplicity in the introduction

[11] $(1/z_g)^{-1} \leq (\langle 1/z_h \rangle)^{-1}$

WCF protocols with arbitrarily small bias, with the least effort.

1.3.1 TDPG-to-Explicit-protocol Framework (TEF) and bias 1/10 | First Contribution

We first describe our framework for converting a TDPG into an explicit protocol. We start by defining a "canonical form" for any given frame of a TDPG. This allows one to write the WCF dual variables, Zs, and the honest state $|\psi\rangle$ associated with each frame of the TDPG. We define a sequence of quantum operations, unitaries and projections, which allow Alice and Bob to transition from the initial frame to the final frame. It turns out that there is only one non-trivial quantum operation in the sequence which we leave partially specified for the moment. This means that we know that the unitary should send the honest initial state to the honest final state. However the action of the unitary on the orthogonal space, which intuitively is what would bestow on it the cheating prevention/detection capability, is obtained as an interesting constraint. Using the SDP formalism we write the constraints at each step of the sequence on the Zs and show that they are indeed satisfied (see Theorem 75 for a full statement of the following, Section 4.2 for its proof and the description of the framework).

Notation. We use $A \geq B$ to mean that $A - B$ has non-negative eigenvalues (we implicitly assume that A and B are hermitian).

Theorem 10 (TEF constraint (simplified))**.** *If a unitary matrix U acting on the space span$\{|g_1\rangle, |g_2\rangle \ldots, |h_1\rangle, |h_2\rangle \ldots\}$ satisfying the constraints*

$$U|v\rangle = |w\rangle,$$

$$\sum_i x_{h_i} |h_i\rangle\langle h_i| - \sum_i x_{g_i} E_h U |g_i\rangle\langle g_i| U^\dagger E_h \geq 0 \tag{1.9}$$

can be found for every transition (see Definition 91 and Definition 25) of a TDPG then an explicit protocol with the corresponding bias can be obtained using the TDPG–to–Explicit–protocol Framework (TEF), where $\{|g_i\rangle\}, \{|h_i\rangle\}$ are orthonormal vectors and if the transition is horizontal

- *the initial points have x_{g_i} as their x-coordinate and p_{g_i} as their corresponding probability weight,*
- *the final points have, similarly, x_{h_i} as their x-coordinate and p_{h_i} as their corresponding probability weight*
- *E_h is a projection onto the span $\{|h_i\rangle\}$ space,*
- *$|v\rangle = \sum_i \sqrt{p_{g_i}} |g_i\rangle / \sqrt{\sum p_{g_i}}$, $|w\rangle = \sum_i \sqrt{p_{h_i}} |h_i\rangle / \sqrt{\sum p_{h_i}}$*

and if the transition is vertical, the x_{g_i} and x_{h_i} become the y-coordinates y_{g_i} and y_{h_i} with everything else unchanged.

Note that the TDPG already specifies the coordinates x_{h_i}, x_{g_i} and the probabilities p_{h_i}, p_{g_i} which satisfy, Equation (1.7), the scalar condition. Our task therefore reduces to finding the correct U which satisfies the aforesaid matrix constraints. It is this general problem that is numerically solved by our EMA algorithm which we describe later.

Given such a unitary U acting on the space span$\{|g_1\rangle, |g_2\rangle \ldots\ |h_1\rangle, |h_2\rangle \ldots\}$ one can construct a unitary, $U^{(2)}_{AM}$, acting non-trivially on the space span$\{|g_1 g_1\rangle_{AM}, |g_2 g_2\rangle_{AM} \ldots, |h_1 h_1\rangle_{AM}, |h_2 h_2\rangle_{AM}, \ldots\}$ by mapping $|g_i\rangle \to |g_i g_i\rangle_{AM}$, $|h_i\rangle \to |h_i h_i\rangle_{AM}$ and as identity otherwise. We now informally describe how to convert a TDPG into an explicit protocol. It suffices to show what a transition from a given frame to the next frame corresponds to in terms of the protocol. In this discussion, we refer

to them as the initial frame and the final frame. Assume that the corresponding non-trivial $U_{AM}^{(2)}$ is known. As we saw, a given transition would either be horizontal or vertical. We assume it is horizontal without loss of generality[12]. We label the points that do not participate in this horizontal transition, i.e. remain unchanged in both frames, by $k_1, k_2 \ldots$ in both frames. The points in the initial frame involved in this transition are labelled $g_1, g_2 \ldots$ and the ones in the final frame are labelled $h_1, h_2 \ldots$. All the points are now labelled. We denote the coordinates of the final points by $x_{h_1}, x_{h_2} \ldots$ and the probability weights by $p_{h_1}, p_{h_2} \ldots$. We similarly define x_{g_i}, p_{g_i} and x_{k_i}, p_{k_i}. The Hilbert space of interest is given by $\mathcal{H} := \text{span}\{|k_1\rangle, |k_2\rangle \ldots, |g_1\rangle, |g_2\rangle \ldots, |h_1\rangle, |h_2\rangle \ldots, |m\rangle\}$ where each vector is assumed orthonormal ($|m\rangle$ is just an idle state in which the message register is assumed to be initially and returned to finally). We assume that Alice's register, Bob's register and the message register each have dimension at least as large as $\dim(\mathcal{H})$. The state (by state in this discussion, we mean the honest state) corresponding to the initial frame is assumed to have the form

$$|\psi_{(1)}\rangle = \left(\sum_i \sqrt{p_{g_i}} |g_i g_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}} |k_i k_i\rangle_{AB} \right) \otimes |m\rangle_M .$$

Bob: Assume Bob has the message register. He applies the conditional swap $U_{BM}^{\text{SWP}\{\vec{g},m\}}$ where $U_{BM}^{\text{SWP}\{\vec{g},m\}}$ swaps conditioned on both registers being in the subspace $\text{span}\{|g_1\rangle, |g_2\rangle \ldots, |m\rangle\}$. The state after this operation is

$$|\psi_{(2)}\rangle = \sum_i \sqrt{p_{g_i}} |g_i g_i\rangle_{AM} \otimes |m\rangle_B + \sum_i \sqrt{p_{k_i}} |k_i k_i\rangle_{AB} \otimes |m\rangle_M .$$

He then sends the message register to Alice.

Alice: Alice applies the non-trivial unitary $U_{AM}^{(2)}$ on her local register and the message register. She then measures $\{E^{(2)}, \mathbb{I} - E^{(2)}\}$ where $E^{(2)} := (\sum |h_i\rangle \langle h_i| + \sum |k_i\rangle \langle k_i|)_A \otimes \mathbb{I}_M$. The state at this point is

$$|\psi_{(3)}\rangle = \sum_i \sqrt{p_{h_i}} |h_i h_i\rangle_{AM} \otimes |m\rangle_B + \sum_i \sqrt{p_{k_i}} |k_i k_i\rangle_{AB} \otimes |m\rangle_M .$$

If the outcome corresponds to the latter, she declares herself to be the winner. Otherwise she sends the message register back to Bob.

Bob: Bob again applies a conditional swap $U_{BM}^{\text{SWP}\{\vec{h},m\}}$ followed by a measurement corresponding to $\{E^{(3)}, \mathbb{I} - E^{(3)}\}$ where $E^{(3)} := (\sum_i |h_i\rangle \langle h_i| + \sum_i |k_i\rangle \langle k_i|)_B \otimes \mathbb{I}_M$. The final state is

$$|\psi_{(4)}\rangle = \left(\sum_i \sqrt{p_{h_i}} |h_i h_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}} |k_i k_i\rangle_{AB} \right) \otimes |m\rangle_M .$$

If the outcome corresponds to $\mathbb{I} - E^{(3)}$, Bob declares himself the winner.

As the final state is in the same form as the initial state, one can progressively build the sequence corresponding to the complete protocol. Once the entire sequence is known, one must reverse the order of all the operations to obtain the final protocol. Note that the message register is initially decoupled, it then gets entangled, and finally it emerges decoupled again. This simplifies the analysis (and also entails that one need not keep the message register coherent for the duration of the protocol; keeping it coherent for each round individually is sufficient).

[12] Mochon's point games have a repeating structure he calls a "ladder". Corresponding to each k he constructs a family of point games parametrised by the number of points in this ladder. The game approaches the bias $\epsilon = (4k+2)^{-1}$ as the number of points is increased (the value is reached in the limit of infinite points). Consequently, we consider a finite set of points in the transition.

Let us try to apply this procedure to our example game (see Figure 1.4). We label the points in the first frame as g_1 and g_2. The state is given by $\frac{1}{\sqrt{2}}\left(|g_1 g_1\rangle_{AB} + |g_2 g_2\rangle_{AB}\right) \otimes |m\rangle_M$. (This should make it clear that the order is reversed here because we want to end with an EPR like state so that when Alice and Bob make a measurement, they agree on a random bit.) We simply claim for the moment that raising does not require Alice and Bob to do anything. This means that we can consider the second frame with the same labels. We now apply the merge transition by using the aforesaid recipe, where Bob applies a swap, sends the message register to Alice, she applies $U^{(2)}_{AM}$ and the projector, returns the message register to Bob and he applies the final swap and measurement. We continue to assume we are given the correct $U^{(2)}_{AM}$ that implements the merge step. The state one obtains after the application of these unitaries turns out to be $|h_1 h_2\rangle_{AB} \otimes |m\rangle_M$. (This looks like the state we should start with, completely unentangled. This is intuitively why the actual protocol is a reversed version of what we have.) Our procedure can be applied to any point game, granted the non-trivial unitary $U^{(2)}$ can be found. The central issue is that no general recipe is known for constructing $U^{(2)}$s.

To address this we can prove that what we call the *Blinkered Unitary* satisfies the required constraints for both the split and merge moves (see Subsection 4.2.1). It is defined as

$$U_{\text{blink}} = |w\rangle\langle v| + |v\rangle\langle w| + \sum_i |v_i\rangle\langle v_i| + \sum_i |w_i\rangle\langle w_i| \tag{1.10}$$

where $|v\rangle$, $\{|v_i\rangle\}$ and $|w\rangle$, $\{|w_i\rangle\}$ are orthonormal vectors spanning the $\{|g_i\rangle\}$ and $\{|h_i\rangle\}$ space respectively. With these the former best protocol (bias $1/6$) can already be derived from its TDPG, in a manner analogous to the one used for the example game. This was not known (to the best of our knowledge), even though the protocol itself was separately known and analysed. We next study the family of bias $1/10$ TDPGs and isolate the precise moves required to implement it (see Subsection 4.3.3). Let $n_g \to n_h$ denote a move from n_g initial points to n_h final points. While the bias $1/6$ games used a $2 \to 1$ merge as its key move, the bias $1/10$ games use a combination of $3 \to 2$ and $2 \to 2$ moves (these can not be produced by a combination of merges and splits, as was pointed out earlier). We give analytic expressions for these unitaries and show that they satisfy the required constraints (see Subsection 4.3.4). In particular, we show that for $3 \to 2$ moves with $x_{g_1} < x_{g_2} < x_{g_3}$ and $x_{h_1} < x_{h_2}$

$$U_{3\to 2} = |w\rangle\langle v| + |w_1\rangle\langle v_1'| + |v_2'\rangle\langle v_2'| + |v_1'\rangle\langle w_1| + |v\rangle\langle w| \tag{1.11}$$

satisfies the required constraints (under some further technical conditions which are satisfied by the $1/10$ games of interest), where

$$\begin{aligned}
|v\rangle &= \frac{\sqrt{p_{g_1}}\,|g_1\rangle + \sqrt{p_{g_2}}\,|g_2\rangle + \sqrt{p_{g_3}}\,|g_3\rangle}{N_g}, \\
|v_1\rangle &= \frac{\sqrt{p_{g_3}}\,|g_2\rangle - \sqrt{p_{g_2}}\,|g_3\rangle}{N_{v_1}}, \\
|v_2\rangle &= \frac{-\frac{(p_{g_2}+p_{g_3})}{\sqrt{p_{g_1}}}\,|g_1\rangle + \sqrt{p_{g_2}}\,|g_2\rangle + \sqrt{p_{g_3}}\,|g_3\rangle}{N_{v_2}}
\end{aligned}$$

and

$$|w\rangle = \frac{\sqrt{p_{h_1}}\,|h_1\rangle + \sqrt{p_{h_2}}\,|h_2\rangle}{N_h}, |w_1\rangle = \frac{\sqrt{p_{h_2}}\,|h_1\rangle - \sqrt{p_{h_1}}\,|h_2\rangle}{N_h}$$

are normalised vectors (this fixes the normalisation factors) which we use to define

$$|v_1'\rangle = \cos\theta\,|v_1\rangle + \sin\theta\,|v_2\rangle\,, |v_2'\rangle = \sin\theta\,|v_1\rangle - \cos\theta\,|v_2\rangle$$

where $\cos\theta$ is obtained by solving

$$\frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}\left(x_{h_1}-x_{h_2}\right)-\cos\theta\frac{\sqrt{p_{g_2}p_{g_3}}}{N_gN_{v_1}}\left(x_{g_2}-x_{g_3}\right)$$
$$-\sin\theta\left\langle x_g\right\rangle\frac{N_g}{N_{v_2}}=0$$

and choosing the solution which is closer to 1. Similarly we give an explicit unitary corresponding to the second move, i.e. the $2\to 2$ move. For the second move, i.e. the $2\to 2$ move with $x_{g_1}<x_{g_2}$ and $x_{h_1}<x_{h_2}$, we show that

$$U_{2\to 2}=|w\rangle\langle v|+\left(\alpha|v\rangle+\beta|w_1\rangle\right)\langle v_1|+|v\rangle\langle w|+\left(\beta|v\rangle-\alpha|w_1\rangle\right)\langle w_1|$$

satisfies the required constraints (again, under further technical conditions which are satisfied by the $1/10$ games of interest) where

$$|v\rangle=\frac{1}{N_g}\left(\sqrt{p_{g_1}}|g_1\rangle+\sqrt{p_{g_2}}|g_2\rangle\right),$$
$$|v_1\rangle=\frac{1}{N_g}\left(\sqrt{p_{g_2}}|g_1\rangle-\sqrt{p_{g_1}}|g_2\rangle\right)$$

and

$$|w\rangle=\frac{1}{N_h}\left(\sqrt{p_{h_1}}|h_1\rangle+\sqrt{p_{h_2}}|h_2\rangle\right)$$
$$|w_1\rangle=\frac{1}{N_h}\left(\sqrt{p_{h_2}}|h_1\rangle-\sqrt{p_{h_1}}|h_2\rangle\right).$$

Further, $\alpha,\beta\in\mathbb{R}$ are such that $\alpha^2+\beta^2=1$ and

$$\beta=\sqrt{\frac{p_{h_1}p_{h_2}}{p_{g_1}p_{g_2}}}\frac{(x_{h_1}-x_{h_2})}{(x_{g_1}-x_{g_2})}.$$

This lets us, in effect, convert Mochon's family of bias $1/10$ games into explicit protocols, finally breaking the $1/6$ barrier. The details of this comprise Chapter 4. Mochon's games achieving lower biases correspond to larger unitary matrices. Consequently, this approach based on guessing the correct form of the solution becomes untenable.

1.3.2 Exact Unitaries for Mochon's assignments | Second Contribution

TEF allows us to convert any TDPG into an explicit protocol, granted the unitaries satisfying Equation (1.9) can be found corresponding to each (valid[13]) transition used in the game (see Theorem 10). Using Kitaev/Mochon's formalism, it is not too hard to see that the following, an even weaker requirement, is enough. Suppose that a valid function can be written as a sum of valid functions. It suffices to find unitaries corresponding to each valid function which appears in the sum[14].

We consider the class of valid functions Mochon uses in his family of point games approaching bias $\frac{1}{4k+2}$ (for arbitrary k). These are of the form (see Definition 80)

$$t=\sum_{i=1}^{n}\frac{-f(x_i)}{\prod_{j\neq i}(x_j-x_i)}[\![x_i]\!]$$

[13] implied by the definition of a TDPG
[14] discussed towards the end of Section 5.2

where $0 \leq x_1 < x_2 \cdots < x_n$ are real numbers, we used Equation (1.8) and $f(x)$ is a polynomial[15]. We refer to these as *f-assignments* and in particular, when f is a monomial, we refer to them as *m-assignments* (in lieu of monomial assignments). We observe that Mochon's f-assignments can be expressed as a sum of m-assignments. We then solve the m-assignments, i.e. give formulae for the unitaries corresponding to m-assignments.

Theorem 11 ((informal[16])). *Let t be Mochon's f-assignment (see Definition 80). Then t can be expressed as $t = \sum_i \alpha_i t'_i$ where α_i are positive and t'_i are monomial assignments. Each t'_i admits an exact solution given in Proposition 87, Proposition 88, Proposition 89, or Proposition 90, depending on the form of t'_i.*

We prove this result in Chapter 5. The first step, is the simple observation that Mochon's f-assignment can be broken into a sum of monomial assignments. The second step involves solving, i.e. finding the unitaries corresponding to, monomial assignments. There are four closely related types of such assignments—balanced/unbalanced aligned/misaligned—whose solutions together with their proof of correctness comprise most of Chapter 5. As an example we state the solution to, what we call, a balanced (aligned) m-assignment. Such an assignment can be written as $t = \sum_{i=1}^{n} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!]$ with $b = m/2$ being an integer. We use $\{|h_1\rangle, |h_2\rangle \ldots |h_n\rangle, |g_1\rangle, |g_2\rangle \ldots |g_n\rangle\}$ as an orthonormal basis. The solution, then, is

$$O = \sum_{i=-b}^{n-b-1} \left(\frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle \langle v'| (X_g)^i \Pi_{g_i}^{\perp}}{\sqrt{c_{h_i} c_{g_i}}} + \text{h.c.} \right)$$

where $X_h := \sum_{i=1}^{n} x_{h_i} |h_i\rangle \langle h_i|, |w\rangle := \sum_{i=1}^{n} \sqrt{p_{h_i}} |h_i\rangle, |w'\rangle := (X_h)^b |w\rangle, c_{h_i} := \langle w'| (X_h)^i \Pi_{h_i}^{\perp} (X_h)^i |w'\rangle$,

$$\Pi_{h_i}^{\perp} := \begin{cases} \text{projector orthogonal to span}\{(X_h)^{-|i|+1} |w'\rangle, (X_h)^{-|i|+2} |w'\rangle \ldots, |w'\rangle\} & i < 0 \\ \text{projector orthogonal to span}\{(X_h)^{-b} |w'\rangle, (X_h)^{-b+1} |w'\rangle, \ldots (X_h)^{i-1} |w'\rangle\} & i > 0 \\ \mathbb{I} & i = 0, \end{cases}$$

and $X_g, |v\rangle, |v'\rangle, c_{g_i}, \Pi_{g_i}^{\perp}$ are defined analogously[17].

Effectively, these constitute the exact description of a family of WCF protocols with bias approaching $1/(4k+2)$, corresponding to Mochon's games.

Relation to the bias $1/10$ protocol. Using this general construction, the unitary corresponding to the key transition/function of the bias $1/10$ game can be constructed, albeit at the cost of increased dimensions (see Example 176). This form is in stark contrast with that of the unitary used in the bias $1/10$ protocol introduced in the previous section. There, the unitary was found perturbatively, obfuscating the underlying general mathematical structure.

There are two shortcomings of this approach. The first is resource requirement. To find the solution to Mochon's f-assignment, we express it as a sum of m-assignments but this costs us an increase in dimensions which in turn corresponds to an increase in the number of qubits required. The second is that it only works for Mochon's family of games. We address these concerns next.

1.3.3 Elliptic Monotone Align (EMA) Algorithm | Third Contribution

We now introduce, what we call, the Elliptic Monotone Align (EMA) algorithm. This algorithm allows one to numerically find the unitary corresponding to any (strictly) valid function. Note that if we neglect the projector in Equation (1.9), we can express it as $X_h \geq U X_g U^\dagger$ where X_h, X_g are diagonal

[15] with some restrictions which we suppress for brevity

[16] We suppressed some constraints on f for brevity.

[17] For the astute reader: by X_h^{-1} we mean $\sum_i x_{h_i}^{-1} |h_i\rangle \langle h_i|$.

matrices with positive entries (justified in Subsection 6.1.1). It is possible to show that we can restrict ourselves to orthogonal matrices without loss of generality (see Subsection 6.1.2). Once we restrict to real numbers, it is easy to see that the set of vectors $\mathcal{E}_{X_h} := \{|u\rangle \mid \langle u| X_h |u\rangle = 1\}$ describe the boundary of an ellipsoid as $\sum_i u_i^2/(x_{h_i}^{-1}) = 1$ (note x_{h_i} is fixed here and u_i is the variable). Similarly $\mathcal{E}_{OX_gO^T}$ represents a rotated ellipsoid where O is orthogonal (see Figure 1.5). Note that larger the x_{h_i} (or x_{g_i}) higher is the curvature of the ellipsoid along the associated direction. It is not hard to see the aforesaid inequality, geometrically, as the $\mathcal{E}_{X_h}$ ellipsoid being contained inside the $\mathcal{E}_{OX_gO^T}$ ellipsoid (the order gets reversed; see Subsection 6.2.1).

Recall that the orthogonal matrix also has the property $O|v\rangle = |w\rangle$. Imagine that in addition, we have $\langle w| X_h |w\rangle = \langle v| X_g |v\rangle$ which in terms of the point game means that the average is preserved (as was the case for merge). In terms of the ellipsoids, it means that the ellipsoids touch along the $|w\rangle$ direction. More precisely, the point $|c\rangle := |w\rangle / \sqrt{\langle w| X_h |w\rangle}$ belongs to both $\mathcal{E}_{X_h}$ and $\mathcal{E}_{OX_gO^T}$. Since the inequality tells us the smaller h ellipsoid is contained inside the larger g ellipsoid, and we now know that they touch at the point $|c\rangle$, we conclude that their normals evaluated at $|c\rangle$ must be equal. Further, we can conclude that the inner ellipsoid must be more curved than the outer ellipsoid.

Mark the point $|c\rangle$ on the $\mathcal{E}_{OX_gO^T}$ ellipsoid. Now imagine rotating the $\mathcal{E}_{X_g}$ ellipsoid to the $\mathcal{E}_{OX_gO^T}$ ellipsoid. The normal at the marked point must be mapped to the normal of $\mathcal{E}_{X_h}$ at $|c\rangle$. It turns out that to evaluate the normals $|n_h\rangle$ on $\mathcal{E}_{X_h}$ at $|c\rangle$ and $|n_g\rangle$ on $\mathcal{E}_{X_g}$ at the marked point, one only needs to know $X_h, X_g, |v\rangle$ and $|w\rangle$. Complete knowledge of O is not required and yet we can be sure that $O|n_g\rangle = |n_h\rangle$ which means O must have a term $|n_h\rangle\langle n_g|$. In fact, one can even evaluate the curvature from the aforesaid quantities. It so turns out that when this condition is expressed precisely, it becomes an instance of the same problem we started with one less dimension allowing us to iteratively find O, which so far we had only assumed to exist. This, however, only works under our assumption that $\langle w| X_h |w\rangle = \langle v| X_g |v\rangle$. This is not always the case. We describe one possible resolution.

Recall that a monotone function f is defined[18] to be a function which has the property "$x \geq y \implies f(x) \geq f(y)$". An operator monotone function[19] is obtained from a generalisation of the aforesaid property to matrices, which in our notation can be expressed as "$X_h \geq OX_gO^T \implies f(X_h) \geq Of(X_g)O^T$". It is known that for a certain class of operator monotone functions f, f^{-1} is also an operator monotone. Using these results in conjunction with results from Aharonov et al. [3] one can show that, after an appropriate scaling of the ellipsoids, there is always an operator monotone f such that $\langle w| f(X_h) |w\rangle = \langle v| f(X_g) |v\rangle$. (This result also admits a simple geometric interpretation. It means that to establish $\mathcal{E}_{X_h}$ is inside $\mathcal{E}_{OX_gO^T}$, which essentially means we look at all different directions and make sure the h ellipsoid is inside the g ellipsoid, we can instead look along a single direction $|w\rangle$ and check that all the different ellipsoids $\mathcal{E}_{f(X_h)}$ are inside the corresponding $\mathcal{E}_{Of(X_g)O^T}$ ellipsoids along just this direction, for every operator monotone f in the class indicated earlier.) Since the orthogonal matrix which solves the initial problem also solves the one mapped by f, we can use our technique on the latter to proceed. It is essentially a combination of these steps that constitutes what we call the Elliptic Monotone Align (EMA) algorithm, which is informally summarised below.

Definition (EMA Algorithm (informal))**.** Given a valid transition (or a transition from a TDPG) the algorithm proceeds in three phases.

1. Initialise
    - Tightening procedure: Bring the final points close to zero until the corresponding ellipsoids start to touch.

[18] not to be confused with the f used for Mochon's f-assignment; they are unrelated

[19] Note that the monotone function $f(x) = x^2$ is not an operator monotone. This is a counter-example: $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \geq \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.

- Spectral domain, matrices: Find the spectrum of the matrices which represent the ellipsoid. Evaluate the smallest matrix size n needed to represent the problem using ellipsoids.
- Bootstrapping: Using the aforesaid, define $\left(X_h^{(n)}, X_g^{(n)}, \left|w^{(n)}\right\rangle, \left|v^{(n)}\right\rangle\right) := \underline{\mathrm{X}}^{(n)}$ where the superscript denotes the size of the matrix and vectors.

2. Iterate (neglecting special cases)

   Input: $\underline{\mathrm{X}}^{(k)}$

   Output: $\underline{\mathrm{X}}^{(k-1)}$, the vector $\left|u_h^{(k)}\right\rangle$ and the orthogonal matrices $\bar{O}_g^{(k)}, \bar{O}_h^{(k)}$

   Procedure:

   - Tightening procedure: Similar to the one above, shrink the outer ellipsoid until it touches the inner ellipsoid.
   - Honest align: Use operator monotone functions to make the ellipsoids touch along the $|w\rangle$ direction.
   - Evaluate the Reverse Weingarten Map: Evaluate the curvatures and the normal (which fixes $\left|u_h^{(k)}\right\rangle$) along the $|w\rangle$ direction.
   - Finite Method: Use the curvatures to specify $\underline{\mathrm{X}}^{(k-1)}$ and find the orthogonal matrices $\bar{O}_g^{(k)}, \bar{O}_h^{(k)}$.

3. Reconstruction

   Evaluate $O^{(n)}$ recursively using $O^{(k)} = \bar{O}_g^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_h^{(k)}$.

Theorem 12 (Correctness of the EMA Algorithm (informal))**.** *Given a transition of a TDPG, the EMA Algorithm always finds a U such that the constraints in Theorem 10 are satisfied.*

See Subsection 1.3.3 for the complete algorithm and proof of the theorem; in particular Definition 140 and Theorem 141 for the corresponding formal statements. Results from a preliminary numerical implementation of the EMA algorithm are discussed in Section 6.4.

Despite the apparent simplicity of the main argument there were many difficulties we had to address in order to prove the aforesaid statement. We had to extend the results about operator monotone functions to be able to use them for performing the tightening step as indicated and for being certain that the solution unitary/orthogonal matrix stays unchanged under these transformations. We also extended some results related to different representations of the aforesaid transitions as these situations arise in the tightening procedure (see Subsection 6.3.2.1). Finding an easy method for evaluating the curvatures—the reverse Weingarten map—was key (see Subsection 6.2.2). The trickiest part of the algorithm, which we have not mentioned here in the introduction, was handling the cases where one of the tangent directions of an ellipsoid has an infinite curvature. For concreteness, imagine an ellipse which under an operator monotone gets mapped to a line segment. The tip of the line segment, if viewed as a limit of an ellipse, has an infinite curvature. In these cases, our finite analysis breaks down as the normal is no longer well defined. For the moves used by Mochon in his $1/18$ game for instance that we tried to numerically solve using this algorithm, this infinite case does not appear. However, to solve the split move using the algorithm (instead of the blinkered unitaries) the infinite case does appear. In either case, our algorithm can handle these infinite cases using what we call the Wiggle-v method (see Subsection 6.3.2.3 and Subsection 6.3.3).

The implication is that we can now numerically convert any point game (including the ones with arbitrarily small bias) into complete protocols. An important remaining question is the effect of noise. In the current analysis two idealising assumptions have been made. First, the EMA algorithm assumes

one can solve certain problems with arbitrary precision classically, such as finding the roots of polynomials and diagonalising matrices. Second, in Mochon/Kitaev's point game formalisms, one assumes that the unitaries are known and applied exactly. Neither of these will hold practically, therefore, the effect of noise on the bias of the protocol must be quantified, which we leave as an open problem for further work.

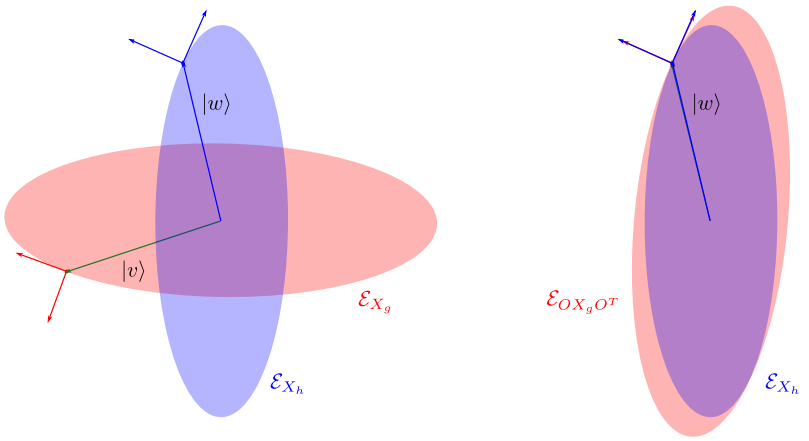


Figure 1.5: On the left the ellipsoids correspond to the diagonal matrices X_g and X_h. The vectors $|w\rangle$ and $|v\rangle$ indicate only the direction. On the right, the larger ellipsoid is now rotated to corresponding to OX_gO^T. The point of contact is along the vector $|w\rangle = O|v\rangle$.

1.3.4 Exact Unitaries for Mochon's assignments, a geometric approach | Fourth Contribution

We give another analytic solution to the monomial assignments based on the ellipsoid picture, allowing us to restate Theorem 11 as follows.

Theorem ((informal[20])). *Let t be Mochon's f-assignment (see Definition 80). Then t can be expressed as $t = \sum_i \alpha_i t'_i$ where α_i are positive and t'_i are monomial assignments. Each t'_i admits an exact solution of the form given in Proposition 165 or Proposition 167.*

We prove this result in Chapter 7 by proceeding as follows. We first show how the unitary corresponding to Mochon's f_0-assignment (where $f(x) = x^0 = 1$) is constructed. This construction has all the basic ingredients needed for constructing the unitaries corresponding to m-assignments and can be explained intuitively.

Consider an f_0-assignment. Let us run the aforementioned argument involving ellipsoids on this assignment, assuming that the contact condition, i.e. $\langle w| X_h |w\rangle = \langle v| X_g |v\rangle$ holds, until the point where we have obtained a problem of the same form with one less dimension. It turns out that for Mochon's f_0-assignment, this contact condition holds. In fact, it turns out that the analogous contact condition continues to hold for all sub-instances of the problem. More precisely, these conditions correspond to $\left\langle (x_h)^k \right\rangle = \left\langle (x_g)^k \right\rangle$ for successively larger integers $k > 0$ (these are satisfied by Mochon's f_0-assignment). Further, since the assignment is a valid function, we know that an O satisfying the necessary conditions exists (see Corollary 102 and Lemma 104). This allows us to iteratively find the solution, O, corresponding to the f_0-assignment (see Section 7.2). It is worth mentioning that the $3 \to 2$ move used for the $1/10$ bias protocol can already be seen as a sum of f_0-assignments.

[20] Again, we suppressed some constraints on f for brevity.

The aforementioned procedure breaks down for m-assignments as the contact condition ceases to hold after a point (after $n-m$ iterations in the simplest case where n = the number of initial points = number of final points). The key idea then, is to use $X_h^{-1} \leq OX_g^{-1}O^T$ instead of $X_h \geq OX_gO^T$. Intuitively, an m-assignment causes the k (as introduced above) to start from m, thereby making it too large after a few iterations, breaking the procedure. Using the inverse leads to a decrease in k and this allows us to find a solution, O, to monomial assignments (see Section 7.6). Together, these also yield WCF protocols with arbitrarily small biases.

For simplicity, we considered here only orthogonal (real unitary) matrices but the actual argument uses isometries and requires the exact expressions for the various geometric quantities (in particular the curvature) to be worked out analytically.

Relation to the EMA algorithm. The EMA algorithm relies on numerical algorithms for diagonalising matrices to reduce the dimension of the problem and for finding solutions to polynomial equations. This stymied the construction of analytic solutions. Here, we remove the need of diagonalising matrices by using three techniques. First, we recast the problem using isometries instead of unitaries, second we derive and use analytic expressions for the various geometric properties that were used, third we restrict ourselves to Mochon's assignment and connect its properties (see [28]) with those appearing geometrically. In the EMA algorithm, the problem of finding solutions to polynomial equations arises as a consequence of alignment using operator monotone functions. Alignment is crucial for reducing the dimension of the problem which is what eventually leads to a solution. Here, given a Mochon's assignment, we show that breaking it into a sum of monomial assignment entails that each monomial assignment in the sum possesses a special property—it is always aligned.

§ 1.4 Navigating the thesis

The thesis may be read linearly. If, however, the reader is interested in reaching the construction of explicit WCF protocols corresponding to Mochon's games at the earliest, and is willing to take some results based on conic duality on faith, then they may consider reading Chapter 2, Chapter 4 followed by Chapter 5. To understand the conic duality see Chapter 3; it is a prerequisite for understanding the EMA algorithm, Chapter 6, which in turn aids the comprehension of Chapter 7. While we don't recommend it, the last chapter may be read directly after Chapter 5.

Colour scheme. In the following chapters, we use purple for intuitive discussions, black for formal statements and blue for proofs.

I

Part

Prior Art

2

Chapter

Existence of (almost perfect) Quantum Weak Coin Flipping Protocols

The contents of this and the following chapter are based on two works: the first is by Carlos Mochon [28] which in turn uses the formalism due to Alexei Kitaev and the second is by Dorit Aharonov, André Chailloux, Maor Ganz, Iordanis Kerenidis, and Loïck Magnin [3] who simplified and verified the former. Essentially all of the informal discussion and intuition is taken from the former and the formal statements (and the accompanying proofs) are taken from the latter.

We briefly review the definition of *coin flipping* before giving a formal definition. The idea is that Alice and Bob are two spatially separated players that do not trust each other. They have a communication channel between them. They wish to agree on a random bit, or so to speak 'flip a coin', by exchanging messages over this channel. They desire a protocol which guarantees that if, say, Alice follows the protocol then Bob can not convince Alice of a biased bit by deviating from the protocol and vice-versa.

Recall that we only consider two kinds of situations, first where both players are honest and second where one player is honest and the other malicious (i.e. deviates from the protocol to increase their winning probability, viz. convincing the honest player that the malicious player has won).

In *weak* coin flipping (WCF), Alice and Bob have opposite preferences and both players are aware of each others preference[1]. We can thus label possible values of the bit as 'Alice wins' and 'Bob wins'.

Coin flipping is sometimes also referred to as *strong* coin flipping (SCF), to emphasise that this assumption about prior knowledge of preferences is not made.

Definition 13 (Weak Coin Flipping)**.** A two party interactive protocol (based on a given model of communication; see below) is such that the initial state is uncorrelated and it ends with both participants outputting a bit. By convention, we assign '0' to mean Alice won, and '1' to mean Bob won. When both parties are honest, i.e. follow the protocol, Alice and Bob output the same bit and their output is uniformly random.

We define P_A^* and P_B^* to be the smallest positive numbers satisfying the following relations. When Alice is honest and Bob deviates from the protocol then the probability of Alice outputting 1 (i.e. Bob won) is less than P_B^*, for all possible deviations. Correspondingly, when Bob is honest and Alice deviates from the protocol then the probability of Bob outputting 0 (i.e. Alice won) is less than P_A^*, for all possible deviations.

We define the bias of the protocol as $\epsilon := \max(P_A^*, P_B^*) - \frac{1}{2}$.

So far we have not explicitly stated the model of communication used by the interacting players. We first consider a (non-relativistic) classical model of communication.

Definition 14 ((Non-relativistic) Classical Communication Model)**.** A *classical model of communication* between two interacting players, Alice and Bob, consists of three parts, Alice's laboratory, Bob's laboratory and a (noiseless) communication channel over which classical messages (or bitstrings without loss of generality) may be shared.

[1] hence if they had the same preference, there would be no need to flip a coin

- An honest player controls only their own laboratory and can interact with the communication channel to send and receive messages. Their laboratory allows them to perform computations on a finite Turing Machine and to save information in a finite memory.
- A cheating player controls everything other than the honest players laboratory. The cheating player (therefore) has access to unbounded computational and storage resources.

Theorem 15. *Any WCF protocol based on a (non-relativistic) classical model of communication allows at least one player to always win if they use their best cheating strategy, i.e. the bias of any such protocol is* $\epsilon = \frac{1}{2}$.

Proof. Restatement of Corollary 9. □

In certain relativistic settings, WCF is possible with non-trivial bias. Here we consider a quantum model of communication (see Definition 16). This and the next chapter are dedicated to proving that under such a communication model, WCF protocols, with arbitrarily small bias, exist.

Are we justified at assuming that the starting state is uncorrelated? Here are the main arguments in support of the assumption.

- If a known entangled state is already shared, then a correlated random bit can be obtained without any communication. The same result can also be obtained by assuming a pre-shared uniformly distributed classical state. However, both methods suffer from the same issue: the players can learn the random bit even before the protocol begins which they can exploit to their advantage.
- The purpose of a single-shot coin flipping protocol should be to create a correlated state thereby preventing players from predicting the outcome before the protocol begins.

It might be possible to weaken the no-correlation requirement to a no-prior-knowledge-of-outcome requirement. Here, however, we do not pursue this line of investigation.

Definition 16 (Quantum Communication Model)**.** A *quantum model of communication* between two interacting players, Alice and Bob, consists of three parts, Alice's laboratory, Bob's laboratory and a (noiseless) communication channel over which quantum states may be exchanged.

- An honest player controls only their own laboratory and can interact with the communication channel to send and receive quantum messages. Their laboratory allows them to perform computations on a quantum computer and to access/initialise a finite number of auxiliary qubits, i.e. arbitrary quantum operations on the incoming state and auxiliary qubits.
- A cheating player controls everything other than the honest players laboratory. The cheating player (therefore) has access to unbounded computational and storage resources which are allowed to be quantum.

The abstract consequences/implicit assumptions of this model are as follows. These clarify the domain of validity of security guarantees.

- Any super-operator encoding the operation performed by a malicious player must act as identity on the honest player's laboratory.
- An honest player's operations are applied as intended.
- An honest player can verify the dimension of the incoming message.

Practically, this might translate, respectively, into the following.

- The laboratory is shielded from electro-magnetic influences to prevent tampering by the malicious player.

- The malicious player has not tampered with the devices of the honest player in advance (so that they function as expected).
- It is hard to verify the incoming system's dimension practically (Mochon laughs it off by saying something to the effect 'a nanobot can't enter through the communication channel'), especially for photonic channels; perhaps with ion traps it is easier.

Notation. We define $\mathbb{R}_{\geq} := [0, \infty)$, $\mathbb{R}_{>} := (0, \infty)$ and similarly $\mathbb{R}_{\leq} := (-\infty, 0], \mathbb{R}_{<} := (-\infty, 0)$. Hermitian conjugate: given a complex number (or matrix) a, by $(a + \text{h.c.})$ we mean $a + a^*$.

§ 2.1 WCF protocol as an SDP and its Dual

Any weak coin flipping protocol can be expressed in the following general form (we do not prove this claim here; see [4]; for an intuitive discussion, see page 9 of [28]).

Definition 17 (WCF protocol with bias ϵ)**.** For n even, an n-message WCF protocol between two players, Alice and Bob, is described by

- three Hilbert spaces with $\mathcal{A}$, $\mathcal{B}$ corresponding to Alice and Bob's private workspaces (Bob does not have any access to $\mathcal{A}$ and Alice to $\mathcal{B}$), and a message space $\mathcal{M}$;
- an initial product state $|\psi_0\rangle = |\psi_{A,0}\rangle \otimes |\psi_{M,0}\rangle \otimes |\psi_{B,0}\rangle \in \mathcal{A} \otimes \mathcal{M} \otimes \mathcal{B}$;
- a set of n unitaries $\{U_1, \dots U_n\}$ acting on $\mathcal{A} \otimes \mathcal{M} \otimes \mathcal{B}$ with $U_i = U_{A,i} \otimes \mathbb{I}_{\mathcal{B}}$ for i odd and $U_i = \mathbb{I}_{\mathcal{A}} \otimes U_{B,i}$ for i even;
- a set of honest states $\{|\psi_i\rangle : i \in [n]\}$ defined by $|\psi_i\rangle = U_i U_{i-1} \dots U_1 |\psi_0\rangle$;
- a set of n projectors $\{E_1, \dots E_n\}$ acting on $\mathcal{A} \otimes \mathcal{M} \otimes \mathcal{B}$ with $E_i = E_{A,i} \otimes \mathbb{I}_{\mathcal{B}}$ for i odd, and $E_i = \mathbb{I}_{\mathcal{A}} \otimes E_{B,i}$ for i even, such that $E_i |\psi_i\rangle = |\psi_i\rangle$;
- two final positive operator valued measures (POVMs) $\{\Pi_A^{(0)}, \Pi_A^{(1)}\}$ acting on $\mathcal{A}$ and $\{\Pi_B^{(0)}, \Pi_B^{(1)}\}$ acting on $\mathcal{B}$.

The WCF protocol proceeds as follows:

- In the beginning, Alice holds $|\psi_{A,0}\rangle |\psi_{M,0}\rangle$ and Bob $|\psi_{B,0}\rangle$.
- For $i = 1$ to n:
  - If i is odd, Alice applies U_i and measures the resulting state with the POVM $\{E_i, \mathbb{I} - E_i\}$. On the first outcome, Alice sends the message qubits to Bob; on the second outcome, she ends the protocol by outputting "0", i.e, Alice declares herself to be the winner.
  - If i is even, Bob applies U_i and measures the resulting state with the POVM $\{E_i, \mathbb{I} - E_i\}$. On the first outcome, Bob sends the message qubits to Alice; on the second outcome, he ends the protocol by outputting "1", i.e., Bob declares himself to be the winner.
  - Alice and Bob measure their part of the state with the final POVM and output the outcome of their measurements. Alice wins on outcome "0" and Bob on outcome "1".

The WCF protocol has the following properties:

- Correctness: When both players are honest, Alice and Bob's outcomes are always the same: $\Pi_A^{(0)} \otimes \mathbb{I}_{\mathcal{M}} \otimes \Pi_B^{(1)} |\psi_n\rangle = \Pi_A^{(1)} \otimes \mathbb{I}_{\mathcal{M}} \otimes \Pi_B^{(0)} |\psi_n\rangle = 0$.
- Balanced: When both players are honest, they win with probability $1/2$: $P_A = \left|\Pi_A^{(0)} \otimes \mathbb{I}_{\mathcal{M}} \otimes \Pi_B^{(0)} |\psi_n\rangle\right|^2 = \frac{1}{2}$ and $P_B = \left|\Pi_A^{(1)} \otimes \mathbb{I}_{\mathcal{M}} \otimes \Pi_B^{(1)} |\psi_n\rangle\right|^2 = \frac{1}{2}$.

- ϵ biased: When Alice is honest, the probability that both players agree on Bob winning is $P_B^* \leq \frac{1}{2} + \epsilon$. And conversely, if Bob is honest, the probability that both players agree on Alice winning is $P_A^* \leq \frac{1}{2} + \epsilon$.

To be able to define the bias, we need to know P_A^* and P_B^* which correspond to the best possible cheating strategy of the opponent. The primal semi-definite program (SDP) formalises this statement.

Theorem 18 (Primal)**.**
$P_B^* = \max Tr((\Pi_A^{(1)} \otimes \mathbb{I}_{\mathcal{M}})\rho_{AM,n})$ *over all* $\rho_{AM,i}$ *satisfying the constraints*

- $Tr_{\mathcal{M}}(\rho_{AM,0}) = Tr_{\mathcal{MB}}(|\psi_0\rangle\langle\psi_0|) = |\psi_{A,0}\rangle\langle\psi_{A,0}|$;
- *for* i *odd,* $Tr_{\mathcal{M}}(\rho_{AM,i}) = Tr_{\mathcal{M}}(E_i U_i \rho_{AM,i-1} U_i^\dagger E_i)$;
- *for* i *even,* $Tr_{\mathcal{M}}(\rho_{AM,i}) = Tr_{\mathcal{M}}(\rho_{AM,i-1})$.

$P_A^* = \max Tr((\mathbb{I}_{\mathcal{M}} \otimes \Pi_B^{(0)})\rho_{MB,n})$ *over all* $\rho_{BM,i}$ *satisfying the constraints*

- $Tr_{\mathcal{M}}(\rho_{MB,0}) = Tr_{\mathcal{AM}}(|\psi_0\rangle\langle\psi_0|) = |\psi_{B,0}\rangle\langle\psi_{B,0}|$;
- *for* i *even,* $Tr_{\mathcal{M}}(\rho_{MB,i}) = Tr_{\mathcal{M}}(E_i U_i \rho_{MB,i-1} U_i^\dagger E_i)$;
- *for* i *odd,* $Tr_{\mathcal{M}}(\rho_{MB,i}) = Tr_{\mathcal{M}}(\rho_{MB,i-1})$.

Proof. We try to express P_B^* as an SDP. To this end, we assume that Bob is malicious while Alice follows the protocol honestly. Consider the (possibly unnormalised) density matrices $\rho_{AM,i}$ for $i \in \{1, 2 \ldots n\}$ defined over the system $\mathcal{AM}$ (see Figure 2.1). The idea is that at the ith step, the density matrix that Alice holds is $\rho_{AM,i-1}$ and after the application of the unitaries (and the projections), it becomes $\rho_{AM,i}$. These density matrices are arbitrary to start with but we constraint them to be consistent with Alice performing the unitaries as described. The remaining freedom in the density matrices corresponds to Bob's cheating strategies. As we maximise over all possible strategies, it is equivalent to maximising over all possible density matrices satisfying the constraints. We now give the constraints. We assume that Bob can't influence Alice's laboratory. Hence, to start with, we must have

$$\mathrm{tr}_M(\rho_{AM,0}) = |\psi_{A,0}\rangle\langle\psi_{A,0}| .$$

Subsequently, we must have

$$\mathrm{tr}_M(\rho_{AM,1}) = \mathrm{tr}_M(E_{A,1} U_{A,1} \rho_{AM,0} U_{A,1}^\dagger E_{A,1})$$

but this needs some explanation. First, let us neglect the projectors. Then, the state after Alice applies her unitary, viz. $U_{A,1}$, the state becomes $U_{A,1}\rho_{AM,0}U_{A,1}^\dagger$. After this, she sends the $\mathcal{M}$ part of the register to Bob which is returned to her after he applies some quantum operation to it (and his local part). This is modelled in two steps. We start by considering an arbitrary state $\rho_{AM,1}$ which must be such that its $\mathcal{A}$ part is the same as that which Alice held, viz. $\mathrm{tr}_M(U_{A,1}\rho_{AM,0}U_{A,1}^\dagger) = \mathrm{tr}_M(\rho_{AM,1})$. We had to add the projectors because Bob immediately loses if Alice's measurement outcome corresponds to $\mathbb{I} - E_{A,1}$. Hence, we only consider the state, weighted by the probability that Alice's measurement corresponded to $E_{A,1}$.

At the next step, again, we only need to ensure that

$$\mathrm{tr}_M(\rho_{AM,2}) = \mathrm{tr}_M(\rho_{AM,1}).$$

In fact, we could have taken $\rho_{AM,2} = \rho_{AM,1}$ without any loss of generality[2] because Alice is not doing anything between these steps, i.e. we could have used $\rho_{AM,2}$ directly instead of $\rho_{AM,1}$. This is useful to

[2] An astute reader might complain that tracing out the message system in $\mathrm{tr}_M(\rho_{AM,1}) = \mathrm{tr}_M(E_{A,1}U_{A,1}\rho_{AM,0}U_{A,1}^\dagger E_{A,1})$ was unwarranted because Alice holds the message register; the constraint should really have been $\rho_{AM,1} = E_{A,1}U_{A,1}\rho_{AM,0}U_{A,1}^\dagger E_{A,1}$. However, the variable $\rho_{AM,1}$ is manifestly redundant. The constraint really is that $\mathrm{tr}_M(\rho_{AM,2}) = \mathrm{tr}_M(E_{A,1}U_{A,1}\rho_{AM,0}U_{A,1}^\dagger E_{A,1})$, i.e. Bob's action on the $\mathcal{M} \otimes \mathcal{B}$ space must leave Alice's part unchanged. It is not hard to see that the set of constraints, as they are written, simplify to essentially the aforesaid. Thus, tracing out M on Alice's turn is justified.

note when we take the dual. The current form, however, is more symmetric because when we analyse for a malicious Alice, $\rho_{BM,1}$ and $\rho_{BM,2}$ would not be equal.

Then, for i odd, we have

$$\mathrm{tr}_M(\rho_{AM,i}) = \mathrm{tr}_M(E_{A,i}U_{A,i}\rho_{AM,i-1}U_{A,i}^\dagger E_{A,i})$$

and for i even, we have

$$\mathrm{tr}_M(\rho_{AM,i}) = \mathrm{tr}_M(\rho_{AM,i-1}).$$

Finally, the probability of a malicious Bob convincing an honest Alice that Bob has won is simply $\mathrm{tr}_{AM}(\rho_{AM,n}\Pi_A^{(1)})$. Proceeding analogously for a malicious Alice and an honest Bob, we obtain the desired result. □

Remark 19. In fact, one can restrict to unitaries without loss of generality (see page 9 of [28]) by simulating the projections as coherent measurements and absorbing them into the final measurement. The projectors, however, can simplify the proofs. Generality is not lost by adding projections because (a) they can only improve the bias and (b) they can be converted into a protocol without projections. One could have, in addition to the measurement $\{E_i, \mathbb{I} - E_i\}$, introduced a similarly defined measurement, say $\{F_i, \mathbb{I} - F_i\}$, before the unitary. This would have yielded $\mathrm{tr}_M(\rho_{AM,i}) = \mathrm{tr}_M(E_iU_iF_i\rho_{AM,i-1}F_iU_i^\dagger E_i)$ for the SDP of P_B^*. We mention this as we use such a form later.

Note that P_B^* depends on Alice's actions (as we optimise over all possible actions of Bob) and thus involves variables such as $\rho_{AM,i}$ and $U_{A,i}$. Analogously, P_A^* depends on Bob's actions.

A feasible solution to an optimisation problem is one which satisfies the constraints but is not necessarily optimal. A feasible solution to the primal problems gives a lower bound on P_A^* and P_B^*. One can instead consider, what is known as, the dual programme. It is known that, a feasible solution to the dual programme gives an upper bound on P_A^* and P_B^*. This certifies how good the protocol is. In fact, in this case, strong duality holds—the optimal value of the dual programme yields P_A^* and P_B^* exactly. In terms of the protocol, it means that there exist cheating strategies corresponding to the optimal values of the dual. We leave the demonstration of strong duality for later and show how to construct the dual programme in the following proof.

Theorem 20 (Dual)**.**
$P_B^* = \min Tr(Z_{A,0}\left|\psi_{A,0}\right\rangle\left\langle\psi_{A,0}\right|)$ *over all* $Z_{A,i}$ *under the constraints*

1. $\forall i,\ Z_{A,i} \geq 0$;
2. *for* i *odd,* $Z_{A,i-1} \otimes \mathbb{I}_{\mathcal{M}} \geq U_{A,i}^\dagger E_{A,i}(Z_{A,i} \otimes \mathbb{I}_{\mathcal{M}})E_{A,i}U_{A,i}$;
3. *for* i *even,* $Z_{A,i-1} = Z_{A,i}$;
4. $Z_{A,n} = \Pi_A^{(1)}$.

$P_A^* = \min Tr(Z_{B,0}\left|\psi_{B,0}\right\rangle\left\langle\psi_{B,0}\right|)$ *over all* $Z_{B,i}$ *under the constraints*

1. $\forall i,\ Z_{B,i} \geq 0$;
2. *for* i *even,* $\mathbb{I}_{\mathcal{M}} \otimes Z_{B,i-1} \geq U_{B,i}^\dagger E_{B,i}(\mathbb{I}_{\mathcal{M}} \otimes Z_{B,i})E_{B,i}U_{B,i}$;
3. *for* i *odd,* $Z_{B,i-1} = Z_{B,i}$;
4. $Z_{B,n} = \prod_B^{(0)}$.

We add one more constraint to the above dual SDPs.

5. $\left|\psi_{A,0}\right\rangle$ *is an eigenvector of* $Z_{A,0}$ *with eigenvalue* $\beta > 0$ *and* $\left|\psi_{B,0}\right\rangle$ *is an eigenvector of* $Z_{B,0}$ *with eigenvalue* $\alpha > 0$.

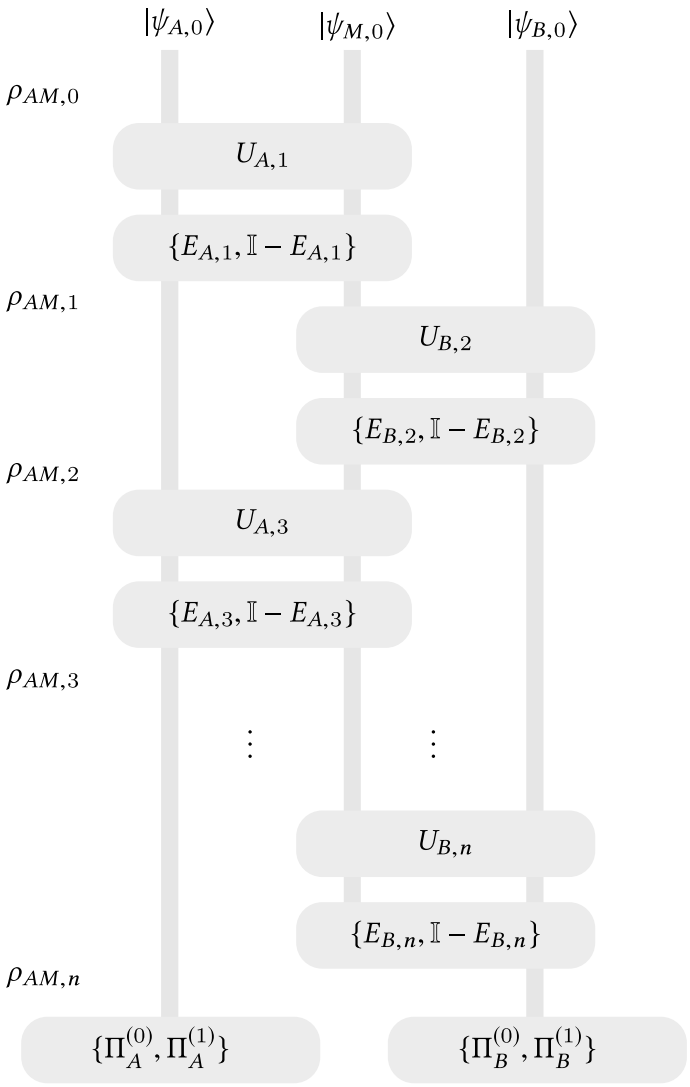


Figure 2.1: Every quantum weak coin flipping protocol can be cast into this general form.

See Appendix A for a proof.

Remark 21. As in Remark 19, the dual SDP for P_B^* would have yielded the constraint

$$Z_{A,i-1} \otimes \mathbb{I}_M \geq F_{A,i} U_{A,i}^\dagger E_{A,i} \left(Z_{A,i} \otimes \mathbb{I}_M\right) E_{A,i} U_{A,i} F_{A,i}$$

for i odd.

Before demonstrating the utility of the 5th constraint (see the next section), we define dual feasible points to be those that satisfy the 5th constraint, or more formally:

Definition 22 (dual feasible points)**.** We call *dual feasible points* any two sets of matrices $\{Z_{A,0}, \ldots, Z_{A,n}\}$ and $\{Z_{B,0}, \ldots, Z_{B,n}\}$ that satisfy the corresponding conditions 1 to 5 as listed in Theorem 20.

The definition is justified by the following proposition.

Proposition 23. *$P_A^* = \inf \alpha$ and $P_B^* = \inf \beta$ where the infimum is over all dual feasible points and β, α are defined in constraint 5 of the definition of the dual feasible points.*

See Appendix A for a proof.

§ 2.2 (Time Dependent) Point Games with EBM transitions/functions

The basic idea here is to remove all inessential information, that is the basis information, from the two aforesaid dual problems. Kitaev's genius was to achieve this by considering, at a given step, the dual variable Z_A, Z_B as observables with $|\psi\rangle$ governing the probability. This combines the evolution of the certificates on cheating probabilities with the evolution of the honest state—the state obtained when both players follow the protocol (nobody cheats). Originally, using a similar manoeuvre, Kitaev settled solvability of the quantum strong coin flipping problem by giving a bound on its bias. To make this insight precise, first "Prob" is defined.

Definition 24 (Prob)**.** Consider $Z \geq 0$ and let $\Pi^{[z]}$ represent the projector on the eigenspace of eigenvalue $z \in \mathrm{sp}(Z)$. We have $Z = \sum_z z\Pi^{[z]}$. Let $|\psi\rangle$ be a (not necessarily normalised) vector. We define the function with finite support $\mathrm{Prob}[Z, \psi] : [0, \infty) \to [0, \infty)$ as

$$\mathrm{Prob}[Z, \psi](z) = \begin{cases} \langle\psi| \Pi^{[z]} |\psi\rangle & \text{if } z \in \mathrm{sp}(Z) \\ 0 & \text{else.} \end{cases}$$

If $Z = Z_A \otimes \mathbb{I}_{\mathcal{M}} \otimes Z_B$, using the same notation, we define the $2-$variate function with finite support $\mathrm{Prob}[Z_A, Z_B, \psi] : [0, \infty) \times [0, \infty) \to [0, \infty)$ as

$$\mathrm{Prob}[Z_A, Z_B, \psi](z_A, z_B) = \begin{cases} \langle\psi| \Pi^{[z_A]} \otimes \mathbb{I}_{\mathcal{M}} \otimes \Pi^{[z_B]} |\psi\rangle & \text{if } (z_A, z_B) \in \mathrm{sp}(Z_A) \times \mathrm{sp}(Z_B), \\ 0 & \text{else.} \end{cases}$$

Think of the aforesaid 2–variate function as assigning a weight on each point of the plane (see, e.g., Figure 1.4). Going from one such configuration to another is what we would intuitively refer to as a "move" for the moment. Notice that at an odd step i, the dual variable Z_B doesn't change while Z_A does (see Theorem 20). The constraint equation in this step is $Z_{A,i-1} \otimes \mathbb{I}_M \geq U_i^\dagger \left(Z_{A,i} \otimes \mathbb{I}_M\right) U_i$, neglecting the projectors. The honest state can be expressed as $|\psi_i\rangle = U_i |\psi_{i-1}\rangle$ but this acts on the complete $\mathcal{A} \otimes \mathcal{M} \otimes \mathcal{B}$ space. Applying the aforesaid method of removing the basis information using the prob method, and appending the fixed $Z_{B,i-1} = Z_{B,i}$, we conclude that $\mathrm{Prob}(Z_{A,i-1} \otimes \mathbb{I}_M \otimes Z_{B,i}, |\psi_{i-1}\rangle) \to \mathrm{Prob}(Z_{A,i} \otimes \mathbb{I}_M \otimes Z_{B,i}, |\psi_i\rangle)$ should constitute an "allowed move" as it is simply re-expressing the dual SDP in a basis independent form. For the dual, we are assuming the protocol is given to us, i.e. U_i (unitary operations), $\Pi_{A/B}$ (measurements) and $|\psi_0\rangle$ (initial state) are specified, and we have to find the appropriate Zs. However, when we discuss the notion of an "allowed move" we are moving towards a framework which will free us from discussing a specific protocol. This motivates the following definitions.

Definition 25 (EBM line transition)**.** Let $g, h : [0, \infty) \to [0, \infty)$ be two functions with finite supports. The line transition $g \to h$ is expressible my matrices (EBM) if there exist two matrices $0 \leq G \leq H$ and a (not necessarily normalised) vector $|\psi\rangle$ such that $g = \mathrm{Prob}\left[G, |\psi\rangle\right]$ and $h = \mathrm{Prob}\left[H, |\psi\rangle\right]$.

Definition 26 (EBM transition)**.** Let $p, q : [0, \infty) \times [0, \infty) \to [0, \infty)$ be two functions with finite supports. The transition $p \to q$ is an

- EBM horizontal transition if for all $y \in [0, \infty)$, $p(., y) \to q(., y)$ is an EBM line transition, and
- EBM vertical transition if for all $x \in [0, \infty)$, $p(x, .) \to q(x, .)$ is an EBM line transition.

Remark 27. When clear from the context, we refer to an EBM line transition also as an EBM transition.

Note that when we wrote the dual, the order of the constraints got inverted, i.e. the condition associated with the final measurements and states appeared first and the condition associated with the initial state appeared in the end. We expect the final state to be like an EPR state and, intuitively, expect two points (in terms of the 2–variate function as described earlier) to be associated with it. This makes it plausible that we will start with two points when the dual is expressed in the aforementioned basis independent way. The initial state of the protocol is unentangled. We expect it to correspond to a single point. This helps us accept that we end with a single point in the basis independent expression of the dual. The rules for moving these points must be related to the dual constraints. We already formalised these conditions into EBM transitions. The notation

$$[\![x_g, y_g]\!](x, y) = \begin{cases} 1 & x_g = x \text{ and } y_g = y \\ 0 & \text{else} \end{cases}$$

will be useful for formalising the complete description into what Mochon dubbed an "Expressible by Matrices" (Time Dependent) point game (TDPG).

Definition 28 (EBM point game)**.** An EBM point game is a sequence of functions $\{p_0, p_1, \ldots, p_n\}$ with finite support such that

- $p_0 = 1/2\,[\![0, 1]\!] + 1/2\,[\![1, 0]\!]$;
- for all even i, $p_i \to p_{i+1}$ is an EBM vertical transition;
- for all odd i, $p_i \to p_{i+1}$ is an EBM horizontal transition;
- $p_n = 1\,[\![\beta, \alpha]\!]$ for some $\alpha, \beta \in [0, 1]$. We call $[\![\beta, \alpha]\!]$ the final point of the EBM point game.

Since we started with a WCF protocol, considered its dual and re-expressed it as a TDPG (which is just a basis independent representation), the following should not come as a surprise.

Proposition 29 (WCF $\implies$ EBM point game)**.** *Given a WCF protocol with cheating probabilities P_A^* and P_B^*, along with a positive real number $\delta > 0$, there exists an EBM point game with final point $[\![P_B^* + \delta, P_A^* + \delta]\!]$.*

See Appendix A for a proof.

What is slightly more non-trivial is that given this EBM (time dependent) point game, or TDPG to be concise, one can construct a WCF protocol. This means that by using only "allowed moves" one can be sure that there exists a corresponding sequence of unitaries U_i, the measurements $\Pi_{A/B}$ and the initial state $|\psi_0\rangle$ complemented by the dual variables $Z_{A,i}$ and $Z_{B,i}$ which certify the bias corresponding to the coordinates of the final point in the point game. This establishes the equivalence between TDPGs and WCF protocols. The precise statement is as follows.

Theorem 30 (EBM to protocol)**.** *Given an EBM point game with final point $[\beta, \alpha]$, there exists a WCF protocol with $P_A^* \leq \alpha$ and $P_B^* \leq \beta$.*

We give a new proof of this theorem (see Chapter 4), which is related to but different from the original proof due to Mochon [28] (and Aharonov et al's version [3]). The key difference is that we place the projector corresponding to the cheat detection step, before the unitary instead of after. This leads to two relative improvements which we outline here for readers familiar with the original proof. The *first improvement* is that in our resultant protocol, the message register gets decoupled from Alice's and Bob's registers after every round while in the original proof yields a protocol where, intuitively, the message register gets more entangled with Alice's and Bob's registers with each round until a point and then gradually gets unentangled with further rounds. The *second improvement* is that the placement of the projector allows us to consider certain matrices with diverging eigenvalues in a well defined way. These pave the path for converting the bias $1/10$ point game (due to Mochon; will be introduced later in this chapter and then with more details in Chapter 6) into a protocol.

§ 2.3 (Time Dependent) Point Games with Valid functions

It is still not easy to check if a given transition is EBM. In this section, we present an alternative characterisation of EBM (line) transitions due to Kitaev/Mochon.

Proposition. *Let $p \to q$ where $p = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ and $q = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$ with all x_{g_i}, x_{h_i} being non-negative and distinct ($x_{g_i} \neq x_{g_j}$ and $x_{h_i} \neq x_{h_j}$ for every $i \neq j$) while p_{g_i}, p_{h_i} are strictly positive. Then the transition is EBM if it is* strictly valid*, i.e. the following equality holds and the inequalities are* strictly *satisfied:*

$$\begin{aligned} \sum_{i=1}^{n_h} p_{h_i} &= \sum_{i=1}^{n_g} p_{g_i} \\ \sum_{i=1}^{n_h} p_{h_i} \frac{\lambda x_{h_i}}{\lambda + x_{h_i}} &\geq \sum_{i=1}^{n_g} p_{g_i} \frac{\lambda x_{g_i}}{\lambda + x_{g_i}} \qquad \forall \quad \lambda > 0 \\ \sum_{i=1}^{n_h} x_{h_i} p_{h_i} &\geq \sum_{i=1}^{n_g} x_{g_i} p_{g_i}. \end{aligned}$$

Conversely, a transition is valid*, i.e. satisfies these inequalities, if the transition $p \to q$ is EBM.*

In fact, the set of EBM functions and the set of valid functions turns out to be the same, up to closures (as alluded to by the definition of strictly valid functions). Its proof uses an interesting connection with operator monotone functions (generalisations of usual monotone functions, as described earlier in Subsection 1.3.3) through conic duality arguments which we defer to the next chapter. This remarkable result removes the matrices from the discussion entirely and trades them for scalar conditions, (albeit infinitely many of those; one for each $\lambda > 0$). The power of this simplification will become apparent shortly.

Notice that whenever p and q have disjoint support, we can equivalently consider the function $t = q - p$ and restate the inequalities above accordingly as

$$\begin{aligned} \sum_{x \in \text{supp}(t)} t(x) &= 0 \\ \sum_{x \in \text{supp}(x)} t(x) f_\lambda(x) &\geq 0 \\ \sum_{x \in \text{supp}(x)} t(x) x &\geq 0 \end{aligned}$$

where

$$f_\lambda(x) = \frac{\lambda x}{(\lambda + x)}. \tag{2.1}$$

We may therefore speak of valid/EBM functions instead of transitions (see Definition 62, Definition 48 and Corollary 65, in the next chapter). One can rephrase Definition 28 in terms of EBM functions instead of transitions, or even valid functions (see Definition 51, in the next chapter). One can also extend the definitions of EBM line/horizontal/vertical transitions to EBM/valid line/horizontal/vertical functions (see Definition 50, in the next chapter).

2.3.1 Examples of valid line transitions

We focus our attention on the simplest cases. In the first case, we ask what can be done with a single point. The second case is that of merging two (or more) points into one. The third is about splitting a single point into two (or more).

Example 31 (Point raise). $p\,[\![x_g]\!] \to p\,[\![x_h]\!]$ with $x_h \geq x_g$ is a valid transition.

Proof. It suffices to show that $f_\lambda(x_h) \geq f_\lambda(x_g)$ for all λ (see Definition 57) when $x_h \geq x_g \geq 0$. However, $f_\lambda(x)$ is a monotonic function.[3] □

Example 32 (Point merge). $p_{g_1}\,[\![x_{g_1}]\!] + p_{g_2}\,[\![x_{g_2}]\!] \to (p_{g_1}+p_{g_2})\,[\![x_h]\!]$ with $x_h \geq \frac{p_{g_1}x_{g_1}+p_{g_2}x_{g_2}}{p_{g_1}+p_{g_2}}$ is a valid transition or more briefly (and generally) $\sum_i p_{g_i}\,[\![x_{g_i}]\!] \to (\sum_i p_{g_i})\,[\![x_h]\!]$ with $x_h \geq \langle x_g \rangle$ is a valid transition.

Proof. We have to show that $(p_{g_1}+p_{g_2})f_\lambda(x_h) \geq p_{g_1}f_\lambda(x_{g_1}) + p_{g_2}f_\lambda(x_{g_2})$ for all $\lambda \geq 0$. It suffices to show this for $x_h = \frac{p_{g_1}x_{g_1}+p_{g_2}x_{g_2}}{p_{g_1}+p_{g_2}}$ (for we can do a point raise for any larger value). In that case, the condition is just a concavity condition on $f_\lambda(x)$ but these monotone functions are indeed concave[4]. As the notion of concavity readily extends to multiple points, the general result easily follows (see Figure 2.2).

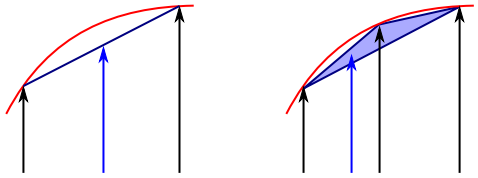


Figure 2.2: The graph illustrates how the notion of concavity readily generalises to multiple points.

□

Example 33 (Point split). $p_g\,[\![x_g]\!] \to p_{h_1}\,[\![x_{h_1}]\!] + p_{h_2}\,[\![x_{h_2}]\!]$ with $p_g = p_{h_1}+p_{h_2}$ and $\frac{p_g}{x_g} \geq \frac{p_{h_1}}{x_{h_1}} + \frac{p_{h_2}}{x_{h_2}}$ is a valid transition or more briefly (and generally) $(\sum_i p_{h_i})\,[\![x_g]\!] \to \sum_i p_{h_i}\,[\![x_{h_i}]\!]$ with $\frac{1}{x_g} \geq \left\langle \frac{1}{x_h} \right\rangle$ is a valid transition.

Proof. We must show that $p_{h_1}f_\lambda(x_{h_1}) + p_{h_2}f_\lambda(x_{h_2}) \geq (p_{h_1}+p_{h_2})\,f_\lambda(x_g)$ for all $\lambda \geq 0$ but it suffices to show it for $f'_\lambda(x) := -\frac{1}{\lambda+x}$ instead of $f_\lambda(x)$ (see the footnote in Example 31). We use $w = \frac{1}{x}$ and assume that the constraint is saturated (if it is not, we can simply do a point raise as before), i.e. $w_g p_g = p_{h_1}w_{h_1} + p_{h_2}w_{h_2}$. We have

$$
\begin{aligned}
(p_{h_1}+p_{h_2})f'_\lambda(x_g) &= (p_{h_1}+p_{h_2})\tilde{f}_\lambda(w_g) && \text{where} \quad \tilde{f}_\lambda(w) := -\frac{w}{1+\lambda w} \\
&\leq (p_{h_1}+p_{h_2})\tilde{f}_\lambda\left(\frac{p_{h_1}w_{h_1}+p_{h_2}w_{h_2}}{p_{h_1}+p_{h_2}}\right) && \because \quad \tilde{f}_\lambda(w) \text{ is monotonically decreasing} \\
&\leq p_{h_1}\tilde{f}_\lambda(w_{h_1}) + p_{h_2}\tilde{f}_\lambda(w_{h_2}) && \because \quad \tilde{f}_\lambda(w) \text{ is convex} \\
&= p_{h_1}f'_\lambda(x_{h_1}) + p_{h_2}f'_\lambda(x_{h_2}).
\end{aligned}
$$

[3] To see this, one possibility is to consider the monotone $f(x) = -x^{-1}$ (for $x \geq 0$) and note that $f'(x) = -(x+\lambda)^{-1}$ will also work (for $\lambda \geq 0$ at least). We can again, add a constant c which would leave the monotonicity unchanged: $f'''(x) = -\frac{1}{x+\lambda} + c = \frac{-1+c(x+\lambda)}{(x+\lambda)} = \frac{\lambda^{-1}x}{x+\lambda}$ if $c = \lambda^{-1}$. Scaling by a positive number also leaves the monotonicity unchanged and hence $f_\lambda(x)$ is a monotone function.

[4] To see this, let's use the form $f(x) = \frac{-1}{x}$ because substituting $x \mapsto x+\lambda$ only shifts the graph to the left, adding an overall constant λ^{-1} shifts the graph upwards, multiplication by a positive constant skews the graph but concavity is clearly preserved under these operations. Finally, to see that $f(x)$ is concave, we note that the second derivative is positive.

The generalisation follows from an argument similar to the one we used in Example 32. □

While these moves appear rather simple, these subsume the earlier protocols in just a few steps which we now discuss.

2.3.2 Example of (Time Dependent) Point Games

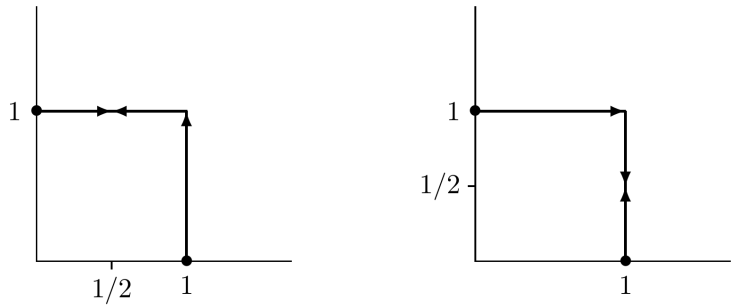


Figure 2.3: Point game for a trivial protocol; bias corresponds to the phone protocol. Taken from [28].

Example 34 (Trivial)**.** A TDPG with bias 1: $\frac{1}{2}[\![0,1]\!] + \frac{1}{2}[\![1,0]\!] \overset{\text{raise}}{\to} \frac{1}{2}[\![1,1]\!] + \frac{1}{2}[\![1,0]\!] \overset{\text{merge}}{\to} [\![1,\frac{1}{2}]\!]$, so $P_A^* \leq \frac{1}{2}$. but $P_B^* \leq 1$ hence the bias is $1/2$ (trivial; see Figure 2.3).

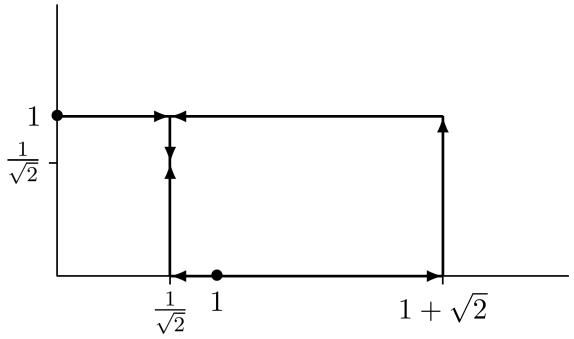


Figure 2.4: Point game for the Spekkens-Rudolph protocol. Taken from [28].

Example 35 (Spekkens-Rudolph (published in Physical Review Letters; see [37]))**.** Fix $x \in (1/2, 1)$, then

$$\begin{aligned}
\frac{1}{2}[\![0,1]\!] + \frac{1}{2}[\![1,0]\!] &\xrightarrow{\text{split}} \frac{2x-1}{2x}[\![x,0]\!] + \frac{1-x}{2x}\left[\!\!\left[\frac{x}{1-x},0\right]\!\!\right] + \frac{1}{2}[\![0,1]\!] \\
&\xrightarrow{\text{raise}} \frac{2x-1}{2x}[\![x,0]\!] + \frac{1-x}{2x}\left[\!\!\left[\frac{x}{1-x},1\right]\!\!\right] + \frac{1}{2}[\![0,1]\!] \\
&\xrightarrow{\text{merge}} \frac{2x-1}{2x}[\![x,0]\!] + \frac{1}{2x}[\![x,1]\!] \\
&\xrightarrow{\text{merge}} 1\left[\!\!\left[x,\frac{1}{2x}\right]\!\!\right].
\end{aligned}$$

Here $P_A^* = \frac{1}{2x}$ and $P_B^* = x$. Setting $P_A^* = P_B^*$ we get $x = 1/\sqrt{2}$ and thus the bias as $\frac{1}{\sqrt{2}} - \frac{1}{2}$ (see Figure 2.4). The first split can be verified to be valid.

Note that the key difference between the trivial point game and the Spekkens-Rudolph point game is the split. This is expected because a raise simply increases the coordinate, and hence the bias, while a merge preserves the average coordinate. However, the points must be aligned for us to be able to merge. While the split increases the average coordinate it is still better than a raise because at least some points are placed below the average coordinate. With one split (in the Spekkens-Rudolph game), we still had to raise but only a part of the weight on the initial point. What happens if we introduce more splits? This turns out to be a good strategy.

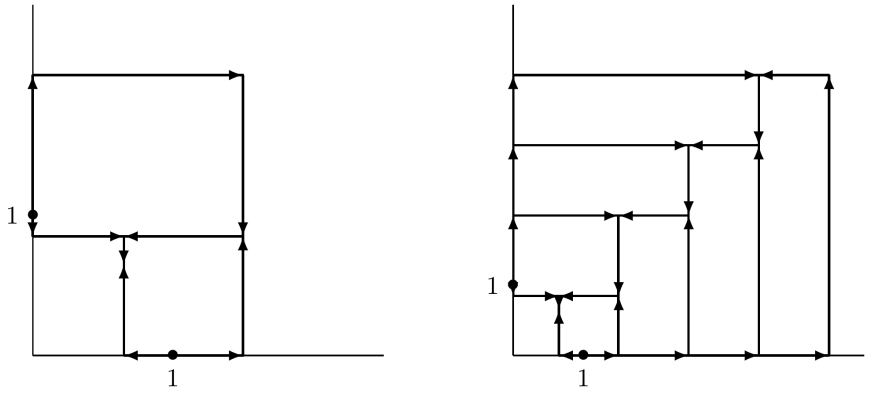


Figure 2.5: Point game for the Dip-dip Boom protocol. Taken from [28].

Example 36 (Dip Dip Boom Protocol (former best known protocol); informal)**.** Suppose we have split the initial points along the axis into so many points that they may effectively be treated using continuous variables to describe the initial frame/configuration as

$$\frac{1}{2}\int_{z^*}^{\infty} p(z)\,[\![z,0]\!]\,dz + \frac{1}{2}\int_{z^*}^{\infty} p(z)\,[\![0,z]\!]\,dz$$

where $p(z)$ is a normalised probability density, i.e. $\int_{z^*}^{\infty} p(z)dz = 1$ (note that z^* is a cutoff; no points are below $[\![0,z^*]\!]$ or to the left of $[\![z^*,0]\!]$). As indicated in Figure 2.5, we want the probability to be taken off of the axes and through the diagonal, into the point $[\![z^*,z^*]\!]$. Suppose the probability[5] at the point $[\![z,z]\!]$ is given by $Q(z)$. To move this point down and right, we can follow a two step procedure: Merge with a point on the x-axis (with weight $\frac{p(z)}{2}dz$) to obtain $(Q(z)+\frac{p(z)}{2}dz)\,[\![z,z-dz]\!]$ and then merge with a point on the y-axis (again, with weight $\frac{p(z)}{2}dz$) to obtain $(Q(z)+p(z)dz)\,[\![z-dz,z-dz]\!]$ which means

$$Q(z-dz) = Q(z) + p(z)dz \implies \frac{dQ}{dz} = p(z).$$

Further, for a merge, we must conserve the average coordinate, i.e. $Q(z)z = \left(Q(z)+\frac{p(z)}{2}dz\right)(z-dz)$ which yields $-Q(z)+\frac{p(z)}{2}z = 0$. We thus have

$$\frac{dQ}{dz} = -\frac{2Q}{z}$$

[5] we really should consider probability densities here but we don't for simplicity; we proceed more carefully in Subsection 4.3.3 of Chapter 4.

which is solved by $Q(z) = c/z^2$ for some c. Since the original distribution was normalised, and the two distributions are linearly related, we can solve for c to get $c = (z^*)^2$. Finally, we enforce that the initial split was valid. This yields $1 \geq \int_{z^*}^{\infty} \frac{p(z)}{z} dz = \int_{z^*}^{\infty} \frac{2(z^*)^2}{z^4} dz = \frac{2}{3z^*}$ and that gives $\frac{2}{3} \leq z^*$. This means we can not construct a protocol with a bias less than $\frac{2}{3} - \frac{1}{2} = \frac{1}{6}$ using this approach. By filling in the details, one can in fact construct a family of point games corresponding to the so-called Dip Dip Boom protocols (for a concise and clear account, see Appendix A of [28]), with bias approaching $1/6$. We also discuss another construction (protocol corresponding to the point game) as a special case of Mochon's games with achieve bias $1/(4k+2)$ when we convert these into explicit protocols (see Subsection 4.3.2).

§ 2.4 Time Independent Point Games (TIPGs)

Mochon's Dip Dip Boom protocol, the one with bias $1/6$, can be expressed already as a (time dependent) point game. However, it is possible to simplify the point game formalism even further and it is in this simplified formalism Mochon constructs his family of point games that achieve an arbitrarily small bias. Instead of worrying about the entire sequence of horizontal and vertical transitions, one can focus on just two functions as described below.

Definition 37 (TIPG)**.** A *time independent point game (TIPG)* is a valid horizontal function a and a valid vertical function b such that

$$a + b = 1 [\![\beta, \alpha]\!] - \frac{1}{2} [\![0, 1]\!] - \frac{1}{2} [\![1, 0]\!]$$

for some $\alpha, \beta > 1/2$. Further

- we call the point $[\![\beta, \alpha]\!]$ the final point of the game, and
- we call the set $\mathcal{S} = ((\text{supp}(a) \cup \text{supp}(b)) \setminus \text{supp}(a + b)$, the set of intermediate points.

Remark 38. When clear from the context, we may use the word TIPG even when $a+b$ is not necessarily $[\![\beta, \alpha]\!] - \frac{1}{2}([\![0, 1]\!] + [\![1, 0]\!])$ but some other function, c, with finite support in $\mathbb{R}_{\geq} \times \mathbb{R}_{\geq}$ satisfying $\sum_{x \in \text{supp}(c)} c(x) = 0$.

Theorem 39 (TIPG to valid point games)**.** *Given a TIPG with a valid horizontal function a and a valid vertical function b such that $a + b = 1[\beta, \alpha] - \frac{1}{2}[0, 1] - \frac{1}{2}[1, 0]$, we can construct, for all $\epsilon > 0$, a valid point game with final point $[\beta + \epsilon, \alpha + \epsilon]$ where the number of transitions depends on ϵ.*

The main difference here is that we do not worry about the sequence in which one must apply the transitions to obtain the final configuration. This justifies the name TIPG which stands for a Time Independent Point Game. It is not too hard to see that if we have a valid point game we can combine the horizontal functions and the vertical functions to obtain a and b. More precisely, if the valid point game with $[\![\beta, \alpha]\!]$ final point is specified by $a_1, a_2 \ldots a_n$ valid horizontal functions and $b_1, b_2 \ldots b_n$ valid vertical functions, then the corresponding TIPG is specified by $a = \sum_{i=1}^{n} a_i$ and $b = \sum_{i=1}^{n} b_i$ which are horizontally and vertically valid respectively and are easily seen to satisfy $a + b = [\![\beta, \alpha]\!] - \frac{1}{2}[\![0, 1]\!] - \frac{1}{2}[\![1, 0]\!]$.

It is a little counter-intuitive in fact to learn that one can convert a TIPG into a valid (time dependent) point game with an arbitrarily small cost on the bias. It is counter-intuitive because it is not clear that one can extract a time ordered sequence as one might, and in fact does for Mochon's point games, run into causal loops that is you expect a point to be present to create another point which in turn is required to produce the first point. The trick that is used to fix this problem is known as the "catalyst

state". One deposits a little bit of weight wherever there is negative weight for a, for instance, and then one can implement a scaled down round of a and b. The scaling is proportional to the weight that is placed to start with. Repeating this procedure multiple times yields the required final state along with the "catalyst state" which stays unchanged. Absorbing the catalyst state leads to a small increase in the bias. The number of rounds increases with how small one wants this increase in bias to be. We give a relevant definition but defer the proof to the Appendix.

Definition 40 (transitively valid)**.** Consider two functions $p, q : \mathbb{R}_{\geq} \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ with finite support. We say the transition $p \to q$ is transitively valid if there exists a sequence of valid transitions $p_0 \to p_1, p_1 \to p_2, \ldots p_{m-1} \to p_m$ such that $p = p_0$ and $q = p_m$.

For the proof of Theorem 39, see Appendix A.

Corollary 41. *Consider a TIPG with a valid horizontal function $a = a^+ - a^-$ and a valid vertical function $b = b^+ - b^-$ such that $a + b = [\![\beta, \alpha]\!] - \frac{1}{2}[\![0, 1]\!] - \frac{1}{2}[\![1, 0]\!]$. Let Γ be the largest coordinate of all the points that appear in the TIPG. Then, for all $\epsilon > 0$, we can construct a point game with $\mathcal{O}\left(\frac{\|b\|\Gamma^2}{\epsilon^2}\right)$ valid transitions and final point $[\![\beta + \epsilon, \alpha + \epsilon]\!]$.*

For a short proof of Corollary 41, see Appendix A.

It is important to state that the conversion from a TIPG to a valid (time dependent) point game, TDPG, is straightforward (if somewhat long) and explicit.

2.4.1 A note on Visualising TDPG and TIPGs

It may be helpful to consider a TDPG and see how it can be represented as a TIPG graphically. To be concrete, consider the trivial game

$$t_1 = \frac{1}{2}\left([\![1, 0]\!] + [\![0, 1]\!]\right) \xrightarrow{\text{vertical}} t_2 = \frac{1}{2}\left([\![1, 1]\!] + [\![0, 1]\!]\right) \xrightarrow{\text{horizontal}} t_3 = \left[\!\left[\frac{1}{2}, 1\right]\!\right].$$

This may be graphically represented using points and arrows (see Figure 2.6). We define $p := t_1$ and $q := t_n$ which in this example is t_3. The first figure represents the transitions $p \to t_2$ and $t_2 \to q$. The corresponding TIPG is given by the horizontally valid function $a = q - t_2$ and the vertically valid function $b = t_2 - p$. The points in the first frame, p, are represented by unfilled squares. The point in the last frame, q, are represented by a filled square. *For the intermediate points* (see Definition 37) it is easy to observe the following: each point in a will be present in b with exactly the opposite weight[6]. In the graph, we follow the convention of representing negative points in b by a unfilled points and positive points by filled points and conversely for points in a—for negative points in a we use filled points and for positive points we use unfilled points. In our example, there is only one *intermediate* point, $[\![1, 1]\!]$ and is represented by a filled point (the last two images in Figure 2.6). For a non-trivial example see Figure 2.7.

Another representation, is to use arrows to essentially convey the same information. The second image in Figure 2.6 and the first image in Figure 2.7 show the connection—the head of a horizontal (vertical) arrow corresponds to a positive (negative) weight for the point in the horizontal (vertical) function.

[6] Recall that for a TIPG $a + b = q - p = [\![\frac{1}{2}, 1]\!] - \frac{1}{2}([\![1, 0]\!] + [\![1, 0]\!])$. This means that all the points in a, excluding those in $\text{supp}(t_3) \cup \text{supp}(t_1)$, must be identical in weight to all the points in b but with exactly the opposite sign, viz. $a(x, y) = -b(x, y)$ for all $(x, y) \in \text{supp}(a) \setminus (\text{supp}(p) \cup \text{supp}(q))$ and $\text{supp}(a) \setminus (\text{supp}(p) \cup \text{supp}(q) = \text{supp}(b) \setminus (\text{supp}(p) \cup \text{supp}(q))$.

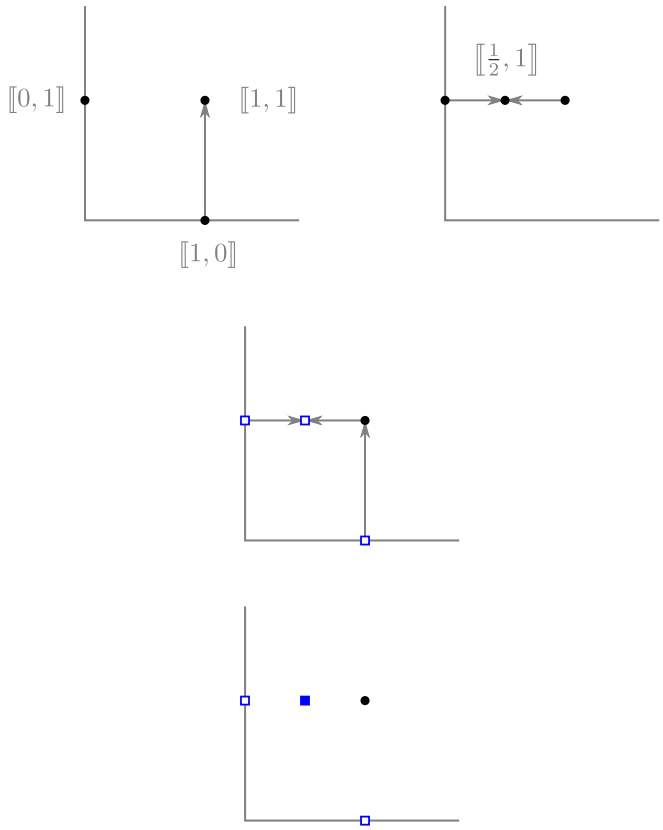


Figure 2.6: Visualising a trivial TDPG point game and its corresponding TIPG (see Subsection 2.4.1).

§ 2.5 Mochon's TIPG achieving bias $\epsilon = 1/(4k+2)$

In this section we show, how we can construct a TIPG for every small $\epsilon > 0$ such that its final point is $[\![\frac{1}{2}+\epsilon, \frac{1}{2}+\epsilon]\!]$, thereby establishing the existence of quantum weak coin flipping protocols with arbitrarily small bias. We begin by giving a brief overview of the construction and subsequently fill in the gaps.

- All the points (excluding $[\![0,1]\!]$ and $[\![1,0]\!]$) involved in these game are placed on a regular lattice, i.e. at locations of the form $[\![a\omega, b\omega]\!]$ where a and b belong to $\mathbb{N}$, and $\omega \in \mathbb{R}_{>}$.
- The family of games is parametrised by k (see Figure 2.8a). These have their final point at $[\![\alpha,\alpha]\!]$ for $\alpha = \zeta\omega = \frac{1}{2} + \mathcal{O}\left(\frac{1}{k}\right)$ where $\zeta \in \mathbb{N}$ and can be calculated (discussed later). This entails that to get a bias ϵ we must have $k = \mathcal{O}(\frac{1}{\epsilon})$.
- These games are perhaps best understood as having three stages (see Figure 2.8b).
  1. *Split.* The point $[\![0,1]\!]$ is vertically split into many points along the y-axis. The resulting points lie between $\zeta\omega$ and $\Gamma\omega$ where ζ and Γ belong to $\mathbb{N}$ (how they are fixed is discussed later). Analogously, the point $[\![1,0]\!]$ is horizontally split into many points along the x-axis.
  2. *Ladder.* Suppose the games are parametrised by k. The ladder consists of points along the diagonal (the second image in Figure 2.8b) and along the axes. The points on the axis are transformed by the ladder into the final points $[\![\alpha - k\omega, \alpha]\!]$ and $[\![\alpha, \alpha - k\omega]\!]$.
  3. *Raise.* The two points $[\![\alpha - k\omega, \alpha]\!]$ and $[\![\alpha, \alpha - k\omega]\!]$ are raised to constitute the final point $[\![\alpha,\alpha]\!]$.

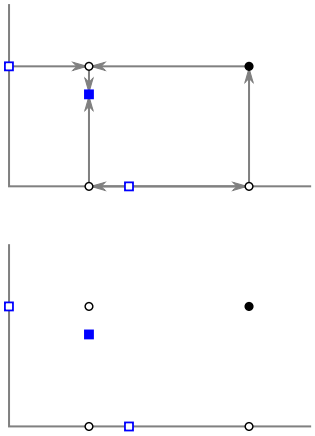


Figure 2.7: Visualising the Spekkens-Rudolph TDPG (above) and TIPG (below).

More concisely, we may describe Mochon's TIPGs as

$$\begin{aligned} \frac{1}{2}[\![0,1]\!]+\frac{1}{2}[\![1,0]\!] &\xrightarrow{\text{split}} \sum_{j=\zeta}^{\Gamma}\text{split}(j)\,[\![0,j\omega]\!]+\sum_{j=\zeta}^{\Gamma}\text{split}(j)\,[\![j\omega,0]\!]\,, \\ &\xrightarrow{\text{ladder}} \frac{1}{2}[\![\alpha-k\omega,\alpha]\!]+\frac{1}{2}[\![\alpha,\alpha-k\omega]\!]\,, \\ &\xrightarrow{\text{raise}} [\![\alpha,\alpha]\!] \end{aligned} \tag{2.2}$$

where the transitions are understood to constitute the final, complete TIPG (see Subsection 2.4.1).

We now show that for each $k \in \{1,2\dots\}$ there exist parameters ω and Γ belonging to $\mathbb{R}_>$ such that (1) the two initial splits are valid, (2) the *ladder* (see Definition 42) corresponds to a horizontally valid and vertically valid function, and (3) $\alpha = \frac{1}{2}+\mathcal{O}\left(\frac{1}{k}\right)$.

2.5.1 The Ladder

Definition 42 (Ladder, Rung)**.** A *ladder* is a TIPG consisting of a valid horizontal function, a_{lad}, and a vertically valid function, b_{lad}, satisfying

$$a_{\text{lad}}+b_{\text{lad}} = \frac{1}{2}[\![\alpha-k\omega,\alpha]\!]+\frac{1}{2}[\![\alpha,\alpha-k\omega]\!]-\sum_{j=\zeta}^{\Gamma}\text{split}(j)\left([\![0,j\omega]\!]+[\![j\omega,0]\!]\right)$$

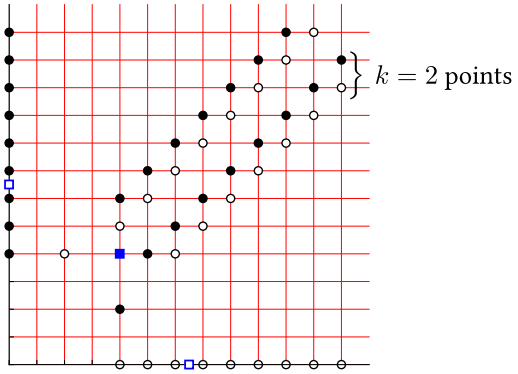
where the parameters Γ, ζ, ω and α are such that for each axis, $\text{split}(j)$ is a distribution on the points arising from splitting the initial points $[\![0,1]\!]$ (or $[\![1,0]\!]$) along the y-axis (or x-axis, respectively), viz. $b_{\text{split}} := \sum_{j=\zeta}^{\Gamma}\text{split}(j)\,[\![0,j\omega]\!]-\frac{1}{2}[\![0,1]\!]$ is a valid vertical function and analogously $a_{\text{split}} := \sum_{j=\zeta}^{\Gamma}\text{split}(j)\,[\![j\omega,0]\!]-\frac{1}{2}[\![1,0]\!]$ is a valid horizontal function.

Further, a *rung* (of a ladder) at a height $j\omega$ is defined to be $a_{\text{rung}} := \sum_{x:(x,j\omega)\in\text{supp}(a)} a_{\text{lad}}(x,j\omega)$.[7]

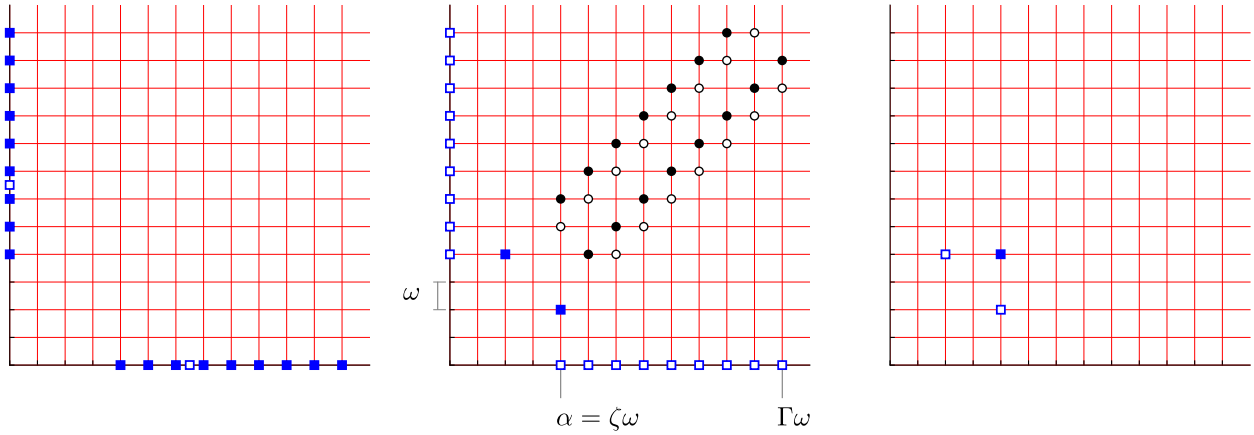
Finally, we use the term *prospective ladder* for a ladder except that its constituent functions a_{lad} and b_{lad} are not necessarily valid.

One way of viewing Mochon's TIPGs is to think of them as generalisations of the bias $1/6$ game but before we do that, we define the notion of symmetric games.

[7] We could have defined a horizontal rung and a vertical rung; however we restrict to symmetric games later making the distinction unnecessary.



(a) Illustration of Mochon's TIPG for $k = 2$.



(b) Mochon's TIPG may be understood in three stages, the initial *splits*, the *ladder*, followed by the *raises*.

Figure 2.8: Mochon's TIPG

Given a ladder (as defined above), we restrict ourselves to *symmetric ladders*, i.e. ladders which satisfy

$$b_{\text{lad}}(x, y) = -b_{\text{lad}}(y, x) \tag{2.3}$$

for each (x, y) belonging to the set of intermediate points (see Definition 37).

Note 43. Given a ladder,

$$a_{\text{lad}}(x, y) = -b_{\text{lad}}(x, y) \tag{2.4}$$

for all (x, y) belonging to the set of intermediate points. (This holds more generally for any TIPG.)

The advantage of restricting to symmetric games is that it halves the work in the sense that we need only establish that a_{lad} is horizontally valid; b_{lad} then is automatically taken care of.

Note 44. Given a symmetric prospective ladder, $a_{\text{lad}}(x, y)$ is horizontally valid if and only if $b_{\text{lad}}(x, y)$ is vertically valid.

This is almost trivial. Fix an arbitrary y and let $f_\lambda(\cdot)$ be an extremal operator monotone parametrised

by λ (see Definition 57). Then, for all $\lambda \geq 0$, we have

$$\begin{aligned}
&\sum_x f_\lambda(x) a_{\text{lad}}(x,y) \geq 0 && \because \quad a_{\text{lad}} \text{ is valid} \\
\iff &-\sum_x f_\lambda(x) a_{\text{lad}}(y,x) \geq 0 && \because \quad (2.4) \\
\iff &\sum_x f_\lambda(x) b_{\text{lad}}(y,x) \geq 0 && \because \quad (2.3) \\
\iff &\sum_x f_\lambda(y) b_{\text{lad}}(y,x) \geq 0 && \because \quad \text{Substitute } x \to y, y \to x
\end{aligned}$$

which means b_{lad} is valid.

The beauty of the point game formalism is perhaps best seen by the fact that it allows one to see successively better protocols as natural generalisations. We already saw this, the trivial protocol could be generalised to the Spekkens Rudolph protocol simply by adding one split. Adding more splits yielded a better bias but we saw that this method saturates at bias $1/6$ (this was the Dip Dip Boom protocol; see Example 36). To go below we use the time independent point game formalism. In this formalism, we will see that Mochon's TIPG are a neat generalisation of the Dip Dip Boom (DDB) game.

To give an intuitive idea, we consider a symmetric variant of the DDB game (without explicitly stating or deriving it). The ladder corresponding to it may be viewed as having exactly one ''line of points'' on either side of the diagonal (see Figure 2.9). In this case, a rung consists of three points and in fact represents a merge (the square point and the filled point have negative weight and the unfilled point has a positive weight). If it is at a height $j\omega$, it has the form $a^{j\omega}_{\text{rung}} = -p_0 [\![0]\!] + p_{(j-1)\omega} [\![(j-1)\omega]\!] - p_{(j+1)\omega} [\![(j+1)\omega]\!]$ where the ps are positive. Increasing the ''density'' of points along the ''lines'', as we saw, only helps us approach bias $1/6$. Mochon's clever insight was to consider a larger number of ''lines'' near the diagonal[8] (see Figure 2.10); he parametrized the number of ''lines'' by $2k$, k on each side of the diagonal. He then went on to show that one can assign weights to these points in such a way that the prospective ladder is indeed a ladder (i.e. the functions a_{lad} (and b_{lad}) are horizontally (and vertically, resp.) valid, given the initial distribution came from a split; see Definition 42).

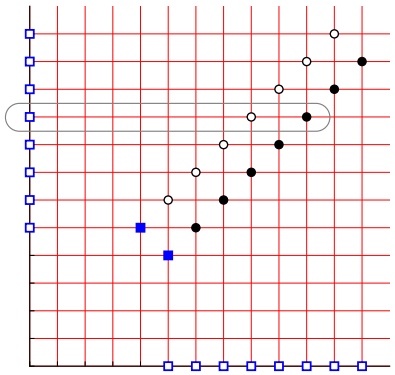


Figure 2.9: The ladder corresponding to a symmetrised version of the Dip Dip Boom protocol.

In Mochon's TIPGs parametrised by k (and ω), a (typical) rung at height $j\omega$, a^j_{rung}, has support only at the axis and near the diagonal, to wit:

$$\text{supp}(a^j_{\text{rung}}) = \{0, (j-k)\omega, (j-(k-1))\omega, \ldots, (j-1)\omega, (j+1)\omega, \ldots (j+(k-1))\omega, (j+k)\omega\}\,.$$

[8] He actually builds his intuition by using the arrow convention and considering the weight circulating inside the loops (see §5.1 Guiding Principles of [28]). Later, Pelchat and Hoyer constructed another family of TIPGs achieving arbitrarily close to zero bias (see [21]).

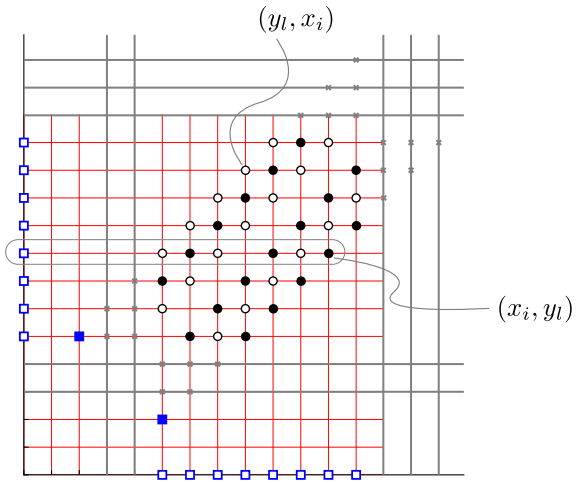


Figure 2.10: Illustration of a ladder corresponding to Mochon's TIPG for $k = 3$.

The key technical tool he introduces is the following: given a set of coordinates, he constructs a way of assigning non-trivial weights to them such that the assignment is valid while still retaining considerable freedom in the assignment. This weight assignment is parametrised by a polynomial and works for essentially all polynomials (up to a given degree). In effect, he further simplifies the validity condition by restricting to a class of functions which are easy to manipulate and are valid by construction.

Lemma 45. *Let*

- $x_1, x_2 \dots x_n$ *be non-negative and distinct real numbers,*
- f *be a polynomial of degree at most* $n-1$ *satisfying* $f(-\lambda) \geq 0$ *for all* $\lambda \geq 0$.

Then, $a = \sum_{i=1}^{n} \frac{-f(x)}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]$ *is a valid function.*

For concreteness, suppose that the coordinates of the points are as illustrated in Figure 2.10. *Ignore* the points on the axis to start with. We would like to assign weights $a_{\text{lad}}(x_i, y_j)$ to each point (x_i, y_j) in the illustration. This we expect to be such that $a_{\text{lad}}(x_i, y_l) = -a_{\text{lad}}(y_l, x_i)$ because we restricted ourselves to symmetric (prospective) ladders (see Equation (2.3)). One way of doing this is to consider a polynomial in two variables, $f(x_i, y_j)$ and use Lemma 45. To be concrete, further consider the set of points highlighted in the illustration and assume they all have their y-coordinate equal to y_l. Suppose we define, with respect to these points, the weights as $a_{\text{lad}}(x_i, y_l) = \frac{-f(x_i, y_l)}{\prod_{j\neq i}(x_j - x_i)}$. Let us focus on the rightmost point (among those highlighted in the illustration). If we pretend that its coordinate is given by (x_i, y_l) then $a_{\text{lad}}(x_i, y_l) = \frac{-f(x_i, y_l)}{1\cdot 2\cdot 4\cdot 5\cdot 6\omega^5}$ because the points are on a lattice of width ω. The corresponding formula for the weight on an arbitrary point (on the lattice) at (y_l, x_i) translates to $a_{\text{lad}}(y_l, x_i) = -\frac{f(y_l, x_i)}{\prod_{j\neq l}(x_j - y_l)}$ where the product is over the set of points with the y-coordinate equal to x_i. If we again pretend that (x_i, y_l) corresponds to the last point on the highlighted set of points in the illustration, then it follows that $a_{\text{lad}}(y_l, x_i) = \frac{-f(y_l, x_i)}{(-1)\cdot(-2)\cdot(-4)\cdot(-5)\cdot(-6)\omega^5}$. This is because (y_l, x_i) now becomes the left most point (and we include the crosses in the counting; why they appear will get clear after Equation (2.5)). When f is a constant function, it is easy to convince oneself that $a_{\text{lad}}(x_i, y_l) = -a_{\text{lad}}(y_l, x_i)$ for other points as well (remember we are ignoring the points on the axis for now). Hence, any symmetric polynomial, i.e. $f(x_i, y_l) = f(y_l, x_i)$, satisfying the requirements of Figure 2.10 in each variable, will yield a valid function satisfying/preserving $a_{\text{lad}}(x_i, y_l) = -a_{\text{lad}}(y_l, x_i)$. However, when we include the points on the axis, we pick an explicit x_i term in the denominator of a_{lad}, i.e. $a_{\text{lad}}(x_i, y_l) = \frac{-f(x_i, y_l)}{x_i 1\cdot 2\cdot 4\cdot 5\cdot 6\omega^5}$ and the argument breaks. To fix it, we replace the

function $f(x_i, y_l)$ by $f'(x_i, y_l) = \frac{1}{y_l} f(x_i, y_l)$ so that $a_{\text{lad}}(x_i, y_l) = \frac{-f'(x_i,y_l)}{x_i 1\cdot 2\cdot 4\cdot 5\cdot 6\omega^5} = -\frac{f(x_i,y_l)}{y_l x_i 1\cdot 2\cdot 4\cdot 5\cdot 6\omega^5}$ and $a_{\text{lad}}(y_l, x_i) = \frac{-f'(y_l,x_i)}{y_l(-1)\cdot(-2)\cdot(-4)\cdot(-5)\cdot(-6)\omega^5} = \frac{-f(y_l,x_i)}{x_i y_l(-1)\cdot(-2)\cdot(-4)\cdot(-5)\cdot(-6)\omega^5}$. Again, the symmetry of the polynomial f, $f(x_i, y_l) = f(y_l, x_i)$, ensures the anti-symmetry of a_{lad}, i.e. $a_{\text{lad}}(x_i, y_l) = -a_{\text{lad}}(y_l, x_i)$ for intermediate points. We now proceed formally.

Let $f(x, y)$ be a symmetric polynomial in the variables x and y, i.e. $f(x, y) = f(y, x)$, which we define shortly. We consider a rung at height $j\omega$, a^j_{rung} and let $\mathcal{S} := \text{supp}(a^j_{\text{rung}})$. We define the rung at height j to be, for some j dependent (but otherwise constant) constant c_j,

$$\begin{aligned} a^j_{\text{rung}} &:= \sum_{x\in\mathcal{S}} \frac{c_j f(x, j\omega)}{\prod_{x'\in\mathcal{S}\setminus\{x\}}(x' - x)} [\![x, j\omega]\!] \\ &= \frac{-c_j f(0, j\omega)}{\prod_{l\in\{-k,\cdots-1,1,\ldots k\}}((j + l)\omega)} [\![0, j\omega]\!] \\ &\quad + \sum_{i\in\{-k,\ldots,-1,1,\ldots k\}} \frac{-c_j f((j + i)\omega, j\omega)}{-(j + i)\omega \prod_{l\in\{-k,\ldots k\}\setminus\{0,i\}}((l - i)\omega)} [\![(j + i)\omega, j\omega]\!] \end{aligned}$$

where the first term is for the point on the axis and the subsequent points are symmetrically distributed across the point $(j\omega, j\omega)$ on the diagonal. As motivated above, we set $c_j = c/j\omega$ and sum over all the rungs to obtain the ladder

$$\begin{aligned} a_{\text{lad}} &= \sum_{j=\zeta}^{\Gamma} a^j_{\text{rung}} \\ &= \sum_{j=\zeta}^{\Gamma} \left(\frac{-c \cdot f(0, j\omega)}{\prod_{l=-k}^{k}((j + l)\omega)} [\![0, j\omega]\!] \right. \\ &\quad \left. + \sum_{i=\{-k,\cdots-1,+1,\ldots k\}} \frac{c \cdot f((j + i)\omega, j\omega)}{((j + i)\omega)(j\omega) \prod_{l\in\{-k,\ldots k\}\setminus\{0,i\}}((l - i)\omega)} [\![(j + i)\omega, j\omega]\!] \right). \end{aligned} \tag{2.5}$$

We now use the freedom in the choice of $f(x, y)$ to remove the points on the edges so that $a_{\text{lad}} + b_{\text{lad}}$ becomes a symmetric ladder (as illustrated in Figure 2.10). Since each rung involves at most $2k + 1$ points, the polynomial can have at most $2k$ degree (in each variable). We define

$$f(x, y) := \underbrace{(-1)^{k+1}}_{\text{I}} \prod_{i=1}^{k-1} [(\alpha - i\omega) - x] \underbrace{[(\alpha - i\omega) - y]}_{\text{II}} \prod_{i=1}^{k} [(\Gamma\omega + i\omega) - x] [(\Gamma\omega + i\omega) - y] . \tag{2.6}$$

Note that for all $\zeta \le j \le \Gamma$,

- $f(x, j\omega)$ has degree at most $2k - 1$
- $f(-\lambda, j\omega) \ge 0$ for all $\lambda \ge 0$.

The second statement holds because there are $k - 1$ terms with a negative sign (term II) which are cancelled by term I (of course, if any one of the term II (in the sum) is zero, the counting becomes irrelevant but the statement holds anyway). All the other terms are non-negative. For each j, we have satisfied the conditions in Lemma 45 and can therefore conclude that a_{lad} is a horizontally valid function. By symmetry, we can also conclude that b_{lad} is a vertically valid function. Formally, we have shown the following.

Lemma 46. *Let the function a_{lad} be as defined in Equation (2.5) and Equation (2.6). Further, let $b_{lad}(x, y) := a_{lad}(y, x)$. Then a_{lad} is a valid horizontal function and b_{lad} is a valid vertical function, satisfying*

$$a_{lad} + b_{lad} = \frac{1}{2} \left([\![\alpha - k\omega, \alpha]\!] + [\![\alpha, \alpha - k\omega]\!] \right) - \sum_{j=\zeta}^{\Gamma} \mathit{split}(j) \left([\![0, j\omega]\!] + [\![j\omega, 0]\!] \right)$$

where

$$\mathit{split}(j) := \frac{c \cdot f(0, j\omega)}{\prod_{l=-k}^{k}((j+l)\omega)}.$$

It remains to prove Lemma 45. It admits an elegant proof and in fact we use this structure later to find the unitaries. However, we defer it to the appendix for the moment (see Section B.3).

2.5.2 The Split (and The Raise)

To establish that a_{lad} and b_{lad} comprise a ladder, we need to show that split(j) arose from a split of the points on the axis. These will put some constraints on Γ but more importantly on ζ which in turn determines the bias.

Lemma 47. *For any k, we can find ω and Γ, such that for $\alpha = \frac{1}{2} + \frac{C}{k}$ (where C is some constant), the functions*

$$a_{\mathit{split}} = \sum_{j=\zeta}^{\Gamma} \mathit{split}(j)\,[\![j\omega, 0]\!] - \frac{1}{2}\,[\![1, 0]\!] \quad \textit{and} \quad b_{\mathit{split}} = \sum_{j=\zeta}^{\Gamma} \mathit{split}(j)\,[\![0, j\omega]\!] - \frac{1}{2}\,[\![0, 1]\!]$$

are valid functions, where $\mathit{split}(j) = \frac{c \cdot f(0,j\omega)}{\prod_{l=-k}^{k}((j+l)\omega)}$ *and* $c = \frac{1}{2} \cdot \left(\sum_{j=\zeta}^{\Gamma} \frac{f(0,j\omega)}{\prod_{l=-k}^{k}((j+l)\omega)}\right)^{-1}$ *where $f(x, y)$ is as defined in Equation* (2.6).

This statement admits a simple but informal demonstration in the limit of $\omega \to 0$ and $\Gamma \to \infty$. We can write, for some k dependent constant c',

$$f(0, z) = c' \left(\prod_{i=1}^{k-1}(\alpha - i\omega - z)\right)\left(\prod_{i=1}^{k}(\Gamma\omega + i\omega - z)\right).$$

In the limit, assuming that $\Gamma - \frac{z}{\epsilon} \to \Gamma$, we can write

$$f(0, z) = c''(z - \alpha)^{k-1}$$

where c'' is some other k dependent constant. As split(j) is the weight on the point $[\![0, z]\!]$ with $z = j\omega$, we write it as $p(z)$. Since we are working in the limit $\omega \to 0$, we have

$$p(z) = \frac{c f(0, z)}{\prod_{l=-k}^{k}(z + l\omega)} \to \frac{c'''(z - \alpha)^{k-1}}{z^{2k+1}}$$

where the constant c''' is fixed by the probability conservation requirement. This requirement, in the integral form, may be expressed as

$$\frac{1}{2} = \int_{\alpha}^{\infty} p(z)dz. \tag{2.7}$$

The split condition may correspondingly be expressed as

$$\frac{1}{2} = \int_{\alpha}^{\infty} \frac{p(z)}{z}dz \tag{2.8}$$

which we saturated to find the best possible value for α. Equating the two, we can cancel c'''. It is useful to note that if we set $z' = \alpha/z$ then we can rewrite the functional form of these as

$$\int_{\alpha}^{\infty} \frac{(z - \alpha)^j}{z^l}dz = \alpha^{j-l+1} \int_{0}^{1} (z')^{l-j-2}(1 - z')^j dz'.$$

The integral is essentially the beta function[9] $B(x, y)$ and therefore we can further write it as being

$$= \alpha^{j-l+1} B(l-j-1, j+1) = \alpha^{j-l+1} \frac{(l-j-2)!j!}{(l-1)!}.$$

Using this result and equating the two constraints (Equation (2.7) and Equation (2.8)), we obtain

$$\cancel{\alpha^{(k-1)-(2k+1)+1}} \frac{\cancel{(2k+1-(k-1)-2)!(k-1)!}}{\cancel{(2k+1-1)!}} = \cancel{\alpha^{(k-1)-(2k+2)+1}} \frac{(2k+2-(k-1)-2)!\cancel{(k-1)!}}{(2k+2-1)!}$$

$$\implies \alpha = \frac{k+1}{2k+1} = \frac{1}{2} + \frac{1}{4k+2}.$$

After the raise, the final point is at $[\![\alpha, \alpha]\!]$ (see Equation (2.2)) which means that the bias is given by $\epsilon = \frac{1}{4k+2}$. For $k = 1$, we recover the bias of the Dip Dip Boom protocol, i.e. $1/6$.

For a formal proof, see Appendix A.

[9] The beta function for x, y is $B(x, y) = \int_0^1 \left(t^{x-1}\right)(1-t)^{y-1}\, dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ where the Gamma function (not to be confused with the constant Γ we used) satisfies $\Gamma(n) = (n-1)!$ for integers $n > 0$. Using $t \to \alpha/t$ we an also write $B(x, y) = \alpha^a \int_0^\infty (t-\alpha)^{b-1} t^{-a-b} dt$.

3

Chapter

Connection with Conic Duality

In Chapter 2, we proved the existence of quantum weak coin flipping protocols with arbitrarily small biases, however, we took a key result, namely the characterisation of EBM transitions/functions on faith. It is perhaps not an overstatement to say that this characterisation is at the heart of weak coin flipping. In this chapter we will see that, the set of EBM functions form a convex cone. The dual to this cone happens to be the set of operator monotone functions (as described earlier in Subsection 1.3.3). These functions have a surprisingly elegant and simple characterisation (a result that is not commonly known to us ordinary mortals, which evidently doesn't include Kitaev). To be able to harness this, one can use the known fact that for a closed convex cone, the dual of the dual is the original cone itself (also called a bi-dual). So this dual of operator monotone functions, i.e. the bi-dual of the cone of EBM functions, equals the cone of EBM functions (up to closures). The dual of operator monotone functions has an easy description because operator monotone functions have an easy description. Combining these, one obtains an easy characterisation of EBM functions. This result was first presented by Mochon/Kitaev but proved using matrix perturbation theory (Appendix C of [28]). The argument we just sketched, however, was also outlined by Mochon/Kitaev. In this chapter, we prove these results following the work of Dorit Aharonov, André Chailloux, Maor Ganz, Iordanis Kerenidis, and Loïck Magnin [3] who worked out a simpler proof along the lines alluded to, by Mochon/Kitaev.

This simplification comes with a catch. We establish that the two cones—the cone of EBM functions and the dual of the cone of operator monotone functions—are the same but given an element in the second, we do not have a recipe for finding the matrices certifying that it is an EBM function (only their existence is guaranteed). Without the matrices, we can not implement the protocol. The remaining chapters are devoted to answering this question to various degrees.

§ 3.1 (Time Dependent) Point Games with valid transitions

3.1.1 Formalising the equivalence between transitions and functions

Working with functions instead of transitions will be rather useful as will be evident from the next subsection.

Definition 48 (K, EBM functions)**.** A function $a : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with finite support is an *EBM function* if the line transition $a^- \to a^+$ is EBM (see Definition 25), where $a^+ : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $a^- : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ denote, respectively, the positive and the negative part of a (i.e. $a = a^+ - a^-$ with $\text{supp}(a^+) \cap \text{supp}(a^-) = \phi$ and $a^{\pm} \geq 0$).

We denote by K the set of EBM functions.

Definition 49 (K_Λ, EBM functions on $[0,\Lambda]$)**.** For any finite Λ, a function $a : [0,\Lambda) \to \mathbb{R}$ with finite support is an *EBM function with support on* $[0,\Lambda]$ if the line transition $a^- \to a^+$ is EBM with its spectrum in $[0,\Lambda]$, where $a^- : [0,\Lambda) \to \mathbb{R}_{\geq 0}$ and $a^+ : [0,\Lambda) \to \mathbb{R}_{\geq 0}$ denote, respectively, the positive and the negative part of a.

We denote the set of EBM functions with support on $[0,\Lambda]$ by K_Λ.

It is evident that if the functions g, h denoting the transition $g \to h$ have no common support, then the function description uniquely captures the said transition. In this section we restrict to such transitions and therefore use them (i.e. functions and transitions) interchangeably. In Chapter 6, we revisit this notion.

To be able to talk about different characterisations of EBM functions, it is useful to abstract it (the characterisation) into a property $\mathcal{P}$ which the function must satisfy. Using this, we can define games which use these $\mathcal{P}$ functions. This facilitates the handling of subtleties which arise in proving that the set of EBM functions is the same as the set of $\mathcal{P}$ functions for specific $\mathcal{P}$s.

Definition 50 (Horizontal and vertical $\mathcal{P}$-functions)**.** A $\mathcal{P}$-function $a : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a function with finite support that has the property $\mathcal{P}$.

A function $t : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a

- *horizontal $\mathcal{P}$-function* if for all $y \geq 0$, $t(.,y)$ is a $\mathcal{P}$-function;
- *vertical $\mathcal{P}$-function* if for all $x \geq 0$, $t(x,.)$ is a $\mathcal{P}$-function.

Suppose $\mathcal{P}$-functions are EBM functions. Consider a TDPG given by the following sequence of valid transitions

$$t_0 = p_0 = \frac{1}{2}\left([\![0,1]\!] + [\![1,0]\!]\right) \to p_1 \to p_2 \cdots \to p_n = [\![\beta,\alpha]\!]$$

and define $t_1 = p_1 - p_0$, $t_2 = p_2 - p_1$, $\ldots t_i = p_i - p_{i-1}$. Clearly, $p_1 = t_1 + t_0$, $p_2 = t_2 + \underbrace{t_1 + t_0}_{p_1}$, $\ldots p_j = \sum_{i=1}^{j} t_i$. This effectively shows how one can construct a TDPG consisting of valid functions, $\{t_i\}$ instead of valid transitions[1], motivating the following definition.

Definition 51 (point games with $\mathcal{P}$-functions)**.** A point game with $\mathcal{P}$-functions is a set $\{t_1,\ldots,t_n\}$ of n $\mathcal{P}$-functions alternatively horizontal and vertical such that

- $\frac{1}{2}[0,1] + \frac{1}{2}[1,0] + \sum_{i=1}^{n} t_i = [\beta,\alpha]$;
- $\forall j \in \{1,\ldots,n\}$, $\frac{1}{2}[0,1] + \frac{1}{2}[1,0] + \sum_{i=1}^{j} t_i \geq 0$.

We call $[\beta,\alpha]$ the final point of the game.

The first condition simply encodes the initial and final frame/configurations while the second condition ensures that the "p_is" are non-negative, i.e. each intermediate frame/configuration is sensible. We already saw that a TDPG with valid transitions can be expressed as a TDPG with valid functions. The reverse also holds.

Lemma 52 ((time dependent) point game with EBM functions $\implies$ (time dependent) point game with EBM transitions)**.** *Given a (time dependent) point game with n EBM functions and final point $[\beta,\alpha]$ we can construct a (time dependent) point game with n EBM transitions and final point $[\beta,\alpha]$.*

[1] We implicitly assumed that the support of p_i and p_{i-1} is disjoint for the non-trivial points (i.e. only the points that participate in the transition). It needs more work to see that even if that is the case, one doesn't lose anything by using functions instead of transitions. We see this in the chapter on the EMA algorithm.

Proof. Suppose the point game with valid functions consists of valid functions $t_1, t_2, \ldots t_n$. We define $p_j := \sum_{i=1}^{j} t_i + \frac{1}{2}\left([\![0,1]\!] + [\![1,0]\!]\right)$. Note that $p_0 = \frac{1}{2}\left([\![0,1]\!] + [\![1,0]\!]\right)$, $p_n = [\![\beta, \alpha]\!]$ and $p_j \geq 0$ for all $j \in \{0, 1 \ldots n\}$, essentially by the definition of a point game with EBM functions. Further,

$$\begin{aligned} p_{j+1} &= p_j + t_{j+1} \\ &= p_j + t^+_{j+1} - t^-_{j+1} \geq 0 \end{aligned}$$

where we split t_{j+1} into its positive and negative part. Since t^+_{j+1} and t^-_{j+1}have disjoint support, we can also conclude that $\zeta = p_j - t^-_{j+1} \geq 0$. One can then rewrite the transition, $p_j \to p_{j+1}$ as

$$\zeta + t^-_{j+1} \to \zeta + t^+_{j+1}.$$

This transition is EBM because $t^-_{j+1} \to t^+_{j+1}$ is EBM which in turn follows from the definition of t_{j+1} being EBM function. □

3.1.2 Operator monotone functions and valid functions

The set of EBM functions forms a convex cone. We recall the definition below.

Definition 53 (convex cone)**.** A set C in a vector space V is a cone if for all $x \in C$ and for all $\lambda > 0$, $\lambda x \in C$. It is convex if for all $x, y \in C$, $x + y \in C$.

Noting that the state $|\psi\rangle$ in the definition of an EBM function (which in turn invokes an EBM transition) is unnormalised, the set of EBM functions is easily seen to form a cone. By taking a direct sum one can establish convexity as well. We state and prove this.

The normed space of interest is the following. Let V be a set of vectors where the vectors are functions from $\mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ with finite support.

- V is an infinite dimensional vector space spanned by $\{[\![x]\!]\}_{x \in \mathbb{R}_{\geq}}$ where $[\![x]\!](y) = \delta_{x,y}$ as defined in the previous chapter. This is evident because we can express each element of V as $v = \sum_x v(x)\,[\![x]\!]$ where the sum is over the finite support[2] of v.
- The usual norm here is the one-norm: $\|v\| := \|v\|_1 = \sum_x |v(x)|$.

Lemma 54. *K is a convex cone. Also, for any $\Lambda \in (0, \infty)$, K_Λ is a convex cone.*

Proof. We give the argument for a finite $\Lambda > 0$ but the reasoning goes through even when the spectrum is unbounded. Let $a, b \in K_\Lambda$ so that $a^- \to a^+$ and $b^- \to b^+$ are two EBM line transitions, viz. we can write them as $a^- = \text{Prob}[X_a, |\psi_a\rangle]$, $a^+ = \text{Prob}[Y_a, |\psi_a\rangle]$, $b^- = \text{Prob}[X_b, |\psi_b\rangle]$ and $b^+ = \text{Prob}[Y_b, |\psi_b\rangle]$. Note that the dimensions of X_a, Y_a may be different from those of X_b, Y_b. To establish that K_Λ is a cone, we have to show that for all $\lambda > 0$, λa is also in the cone. To see this, note that $\lambda a = \lambda a^+ + \lambda a^- = \text{Prob}[Y_a, \sqrt{\lambda}\,|\psi_a\rangle] - \text{Prob}[X_a, \sqrt{\lambda}\,|\psi_a\rangle]$ and so $\lambda a^- \to \lambda a^+$ is EBM with spectra in $[0, \Lambda]$ and hence in K_Λ.

We now show that K_Λ is also convex. It suffices to show that $a + b \in K_\Lambda$. It is easy to see that $a^- + b^- = \text{Prob}[X, |\psi\rangle]$ and $a^+ + b^+ = \text{Prob}[Y, |\psi\rangle]$ if we define

$$X = X_a \oplus X_b = \begin{bmatrix} X_a & \\ & X_b \end{bmatrix}, \quad Y = Y_a \oplus Y_b = \begin{bmatrix} Y_a & \\ & Y_b \end{bmatrix}$$

[2] If instead of a sum, we had used an integral there, we would have had to use a Dirac delta function/distribution. However, restricting to finitely supported functions suffices for our purpose.

and $|\psi\rangle = |\psi_a\rangle \oplus |\psi_b\rangle = \begin{bmatrix} |\psi_a\rangle \\ |\psi_b\rangle \end{bmatrix}$. Further, the matrices have have their spectra contained in $[0, \Lambda]$, i.e. $\text{spec}(X), \text{spec}(Y) \subseteq [0, \Lambda]$. As $Y \geq X$ follows directly from $Y_{a/b} \geq Y_{a/b}$, we conclude $a + b \in K_\Lambda$ as asserted. □

To establish an alternative characterisation of the cone of EBM functions, we recall the definition of a dual cone.

Definition 55 (dual cone)**.** Let C be a cone in a normed vector space V. We denote by V' the space of continuous linear functionals from V to $\mathbb{R}$. The dual cone of a set $C \subseteq V$ is

$$C^* = \left\{\Phi \in V' | \forall a \in C, \Phi(a) \geq 0\right\}.$$

For our purpose, linear functionals can be thought of simply as functions which map objects in the cone to a non-negative real number with the added property of being linear in its argument. We now formally define operator monotone functions.

Definition 56 (operator monotone functions)**.** A function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is operator monotone if for all $0 \leq X \leq Y$ we have $f(X) \leq f(Y)$.

Definition 57 (operator monotone functions on $[0, \Lambda]$)**.** A function $f : [0, \Lambda] \to \mathbb{R}$ is operator monotone on $[0, \Lambda]$ if for all $0 \leq X \leq Y$ with spectrum in $[0, \Lambda]$ we have $f(X) \leq f(Y)$.

The pivotal result here is the equivalence between the cone of operator monotone functions and the dual cone of EBM functions. To state this formally, we consider the following isomorphism. There is a bijective mapping between $\Phi \in V'$ (the space of linear functionals; see Definition 55) and f_Φ which is defined as $f_\Phi(x) = \Phi([\![x]\!])$. Thus, by linearity, for any $h = \sum_x h(x) [\![x]\!]$ we have $\Phi(\sum_x h(x) [\![x]\!]) = \sum_x h(x) f_\Phi(x)$. We can therefore see elements of the dual cone as functions on real numbers (but these functions themselves needn't be linear). We can now state and prove the result. The proof follows from the respective definitions after appropriate unpacking.

Lemma 58. *$\Phi \in K^*$, the dual to the set of EBM functions, if and only if f_Φ is operator monotone in $[0, \infty]$. Also, for any $\Lambda \in (0, \infty)$, $\Phi \in K^*_\Lambda$ if and only if f_Φ is operator monotone on $[0, \Lambda]$.*

Proof. Again, we prove the result for $\Lambda > 0$ but the reasoning holds for K^* as well (and not just for K^*_Λ). Fix any $\Lambda > 0$.

We start with the forward implication: $\Phi \in K^*_\Lambda \implies f_\Phi$ is operator monotone in $[0, \Lambda]$. Recall, $\Phi \in K^*_\Lambda$ implies that

$$\sum_x f_\Phi(x) h(x) \geq 0 \tag{3.1}$$

for all $h \in K_\Lambda$. First, note that

$$\begin{aligned} \sum_x f_\Phi(x) \text{Prob}[X, |\psi\rangle](x) &= \sum_x f_\Phi(x) \langle\psi| \Pi_X^{[x]} |\psi\rangle \\ &= \langle\psi| \sum_x f_\Phi(x) \Pi_X^{[x]} |\psi\rangle \\ &= \langle\psi| f_\Phi(X) |\psi\rangle . \end{aligned} \tag{3.2}$$

Now, observe the following equivalences:

$$\begin{aligned}
&\forall h \in K_\Lambda, \quad \sum_x f_\Phi(x)h(x) \geq 0 \\
\iff &\forall \left|\psi\right\rangle, \quad \forall\, 0 \leq X \leq Y \text{ with } \mathrm{spec}(X), \mathrm{spec}(Y) \subseteq [0,\Lambda], \\
&\sum_x f_\Phi(x)\left(\mathrm{Prob}[Y,\left|\psi\right\rangle](x) - \mathrm{Prob}[X,\left|\psi\right\rangle](x)\right) \geq 0 \\
\iff &\forall \left|\psi\right\rangle, \quad \forall \quad 0 \leq X \leq Y \text{ with } \mathrm{spec}(X), \mathrm{spec}(Y) \subseteq [0,\Lambda], \\
&\left\langle\psi\right| f_\Phi(X) \left|\psi\right\rangle \leq \left\langle\psi\right| f_\Phi(Y) \left|\psi\right\rangle \qquad \text{using Eq. (3.2)} \\
\iff &f_\Phi \text{ is operator monotone on } [0,\Lambda].
\end{aligned}$$

The backwards implication also almost follows from the aforesaid equivalences. We still need to prove that Φ is continuous[3]. Since f_Φ is an operator monotone function on $[0,\Lambda]$, f_Φ is increasing and therefore for all $x \in [0,\Lambda]$ we have $f_\Phi(x) \in [f_\Phi(0), f_\Phi(\Lambda)]$ which entails $\|f_\Phi\|_\infty < \infty$ (the ∞ norm essentially picks the largest element from the set $\{|f_\Phi(x)| : x \in [0,\Lambda]\}$, in this case, either $f_\Phi(0)$ or $f_\Phi(\Lambda)$). Thus, for any $h = \sum_x h(x)\,[\![x]\!]$, we have [4]

$$\Phi(h) = \sum_x h(x) f_\Phi(x) \leq \|h\|_1 \|f_\Phi\|_\infty .$$

which establishes that $\Phi(h)$ is continuous (because this implies that $\lim_{\delta h \to 0} \Phi(\delta h) = 0$ and by linearity, $\lim_{\delta h \to 0} [\Phi(h) - \Phi(h+\delta h)] = \lim_{\delta h \to 0} \Phi(\delta h) = 0$). □

What makes this connection interesting is the following beautiful characterisation of operator monotone functions introduced by Löwner (in 1934, see [8]) which we state shortly without attempting a proof. However, it is not too hard to at least partially motivate it by proving simpler statements. The functions $f(x) = x$ and $f(x) = 1$ are trivially operator monotone. We already saw in the first chapter that the function $f(x) = x^2$ is monotone (on $[0,\infty)$) but it is not operator monotone. (However, all operator monotone functions are monotone.) We now show that $f(x) = -x^{-1}$ is an operator monotone (on $(0,\infty)$). To see this, we start with $0 < X \leq Y$. We have

$$\begin{aligned}
& X \leq Y \\
\iff \quad & X^{-1/2} X X^{-1/2} \leq X^{-1/2} Y X^{-1/2} \\
\iff \quad & \mathbb{I} \leq X^{-1/2} Y X^{-1/2} \\
\iff \quad & \mathbb{I} \geq (X^{-1/2} Y X^{-1/2})^{-1} \\
\iff \quad & \mathbb{I} \geq X^{1/2} Y^{-1} X^{1/2} \\
\iff \quad & X^{-1} \geq Y^{-1} \\
\iff \quad & -X^{-1} \leq -Y^{-1}
\end{aligned}$$

where the first equivalence holds because conjugation with a full rank symmetric matrix (all matrices here are assumed symmetric) doesn't affect the inequality[5], the second is trivial, the third holds because symmetric matrices can be diagonalised and the inequality only states that the eigenvalues of $X^{-1/2} Y X^{-1/2}$ are greater than 1, so their inverses are less than 1, the fourth is trivial and in the fifth we conjugated again.

Thus far we have found the following operator monotone functions: $f(x) = 1, x, -x^{-1}$. It is not hard to see that if we start with $X \leq Y$ and "shift the origin" by adding a constant matrix, $X + \lambda\mathbb{I} \leq Y + \lambda\mathbb{I}$

[3] This is because Φ was supposed to be a *continuous* linear functional; see Definition 55.

[4] we use $\sum_x h(x) f_\Phi(x) \leq \sum_x |h(x)|\,|f_\Phi(x)| \leq \|f_\Phi\|_\infty \sum_x |h(x)| \leq \|f_\Phi\|_\infty \|h\|_1$

[5] ($X \leq Y \iff \forall \left|v\right\rangle, \left\langle v\right| X \left|v\right\rangle \leq \left\langle v\right| Y \left|v\right\rangle \iff \forall M \left|v\right\rangle, \left\langle v\right| M^\dagger X M \left|v\right\rangle \leq \left\langle v\right| M^\dagger Y M \left|v\right\rangle \iff M^\dagger X M \leq M^\dagger Y M$)

and subsequently apply, say $f(x) = -x^{-1}$, the reasoning should go through[6]. Thus, we conclude that $f(x) = \frac{-1}{\lambda+x}$ is an operator monotone. Applying the ''shift of origin'' again, but this time by λ^{-1}, we obtain $f(x) = \frac{-1}{\lambda+x} + \lambda^{-1} = \frac{\lambda^{-1}x}{\lambda+x}$ is also an operator monotone. Finally, scaling by a positive constant such as λ^2 changes nothing, and so conclude that $f_\lambda(x) := \frac{\lambda x}{\lambda+x}$ is also an operator monotone (we did this so that the function $f_\lambda : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$). The striking fact is that these functions, 1, x and $f_\lambda(x)$ are the extremal rays of the cone of operator monotone functions and together span the entire cone.

Lemma 59 (characterisation of operator monotone functions [8])**.** *Any operator monotone function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ can be written as*

$$f(x) = c_0 + c_1 x + \int_0^\infty \frac{\lambda x}{\lambda + x} d\omega(\lambda)$$

for a measure ω satisfying $\int_0^\infty \frac{\lambda}{1+\lambda} d\omega(\lambda) < \infty$.

Lemma 60 (characterisation of operator monotone functions on $[0, \Lambda]$ [8])**.** *Any operator monotone function $f : [0, \Lambda] \to \mathbb{R}$ can be written as*

$$f(x) = c_0 + c_1 x + \int \frac{\lambda x}{\lambda + x} d\omega(\lambda)$$

with the integral ranging over $\lambda \in (-\infty, -\Lambda) \cup (0, \infty)$ satisfying $\int \frac{\lambda}{1+\lambda} d\omega(\lambda) < \infty$ where ω is a measure.

As will become clear when we discuss the dual of the cone of operator monotones, it suffices to consider the extremal rays of the cone, i.e. operator monotones of the form $\lambda x/(\lambda + x)$ (together with 1 and x).

It is known that the bi-dual of a cone is the closure of the cone we started with.

Fact 61. *Let $C \subseteq V$ be a convex cone, then $C^{**} = cl(C)$ where C^* is the dual cone of C.*[7]

The astute reader would have guessed where we are going with this discussion. We define, from hindsight, the bi-dual of EBM functions to be the cone of valid functions. Since the dual of EBM functions has an easy characterisation, the bi-dual also has an easy characterisation which is why we are interested in it.

Definition 62 (Λ-valid functions)**.** A function $a : [0, \Lambda] \to \mathbb{R}$ with finite support on $[0, \Lambda]$ is Λ-valid if $a \in K_\Lambda^{**}$.

To be able to use the aforementioned fact we note that the cone of interest, the cone of EBM functions, is closed when the matrices involved have a bounded spectrum. In this case, it means that the cone of valid functions is the same as the cone of EBM functions. We state this precisely below.

Lemma 63. *For $\Lambda \in (0, \infty)$, K_Λ is closed (which implies $K_\Lambda^{**} = K_\Lambda$).*

Proof. Fix any $\Lambda > 0$ and consider a converging sequence of functions, $\{t_i\}_{i \in \mathbb{N}}$, in the cone K_Λ. Let the limit of this sequence be $t = \lim_{i \to \infty} t_i$. It suffices to show that t is also in K_Λ. We denote t as $t = \sum_{x \in S} t(x) [\![x]\!]$ where S is the support of t. Since t is an element of V (defined after Definition 55), it

[6] with some minor changes to the interval over which the function remains operator monotone

[7] See [Boyd and Vandenberghe 2004] for proofs of these facts.

has finite support. We write $t_i = \text{Prob}[Y_i, |\psi_i\rangle] - \text{Prob}[X_i, |\psi_i\rangle]$ for some $Y_i \geq X_i$ with their spectrum in $[0, \Lambda]$, as $t_i \in K_\Lambda$. We write their spectral decomposition as

$$X_i = \sum_{x \in \text{spec}(X_i)} x \Pi_{X_i}^{[x]}, \quad \text{and} \quad Y_i = \sum_{x \in \text{spec}(Y_i)} y \Pi_{Y_i}^{[y]},$$

where $\Pi_{X_i}^{[x]}$ is a projector onto the eigenspace of X_i with eigenvalue x. We define

$$A_i = \sum_{x \in S} x \Pi_{X_i}^{[x]}, \quad \text{and} \quad B_i = \sum_{y \in S} y \Pi_{Y_i}^{[y]} + \sum_{y \in \text{spec}(Y_i) \setminus S} \Lambda \cdot \Pi_{Y_i}^{[y]},$$

where we have essentially replaced all the eigenvalues of X, not in S, with zeros and those of Y with Λs. Consequently, we have

$$0 \leq A_i \leq X_i \leq Y_i \leq B_i.$$

Using these, we define the EBM functions $t'_i = \text{Prob}[B_i, |\psi_i\rangle] - \text{Prob}[A_i, |\psi_i\rangle]$. One can assume, without loss of generality (see Lemma 104 or Lemma 43 in Mochon's work [28]), that all the matrices $\{A_i\}$ and $\{B_i\}$ have size $s \times s$ and the vector $|\psi\rangle$ is also of size s, with $s = 2|S|$. We express t_i as $t_i = u_i + v_i$ with

$$u_i = \sum_{x \in S} t_i(x) [\![x]\!] \quad \text{and} \quad v_i = \sum_{x \in \text{supp}(t_i) \setminus S} t_i(x) [\![x]\!].$$

Further, let $\epsilon_i = \sum_{x \in \text{supp}(t_i) \setminus S} t_i(x)$ and note that since $\lim_{i \to \infty} t_i = t$, we must have $\lim_{i \to \infty} \epsilon_i = 0$. We define $t'_i := u_i + \epsilon_i^+ [\![\Lambda]\!] - \epsilon_i^- [\![0]\!]$ where $\epsilon_i^\pm$ are the positive/negative parts of ϵ_i. It follows from our construction that $t' \in K_\Lambda$ because $t'_i = \text{Prob}[B_i, |\psi_i\rangle] - \text{Prob}[A_i, |\psi_i\rangle]$. Noting that ϵ_i^+ and ϵ_i^- essentially capture the weight on the points outside spec(t), it follows that $\|t'_i - t_i\|_1 \leq 2\epsilon_i$. It is then clear that $\lim_{i \to \infty} t'_i = t$ since $\lim_{i \to \infty} \epsilon_i = 0$ and $\lim_{i \to \infty} t_i = t$.

We now show that $\lim_{i \to \infty} t'_i$ belongs to the cone, K_Λ. Let X_Λ^s denote the set of $s \times s$ positive semi-definite matrices with their spectra in $[0, \Lambda]$ and Y^s denote the set of quantum states of dimension s. Consider the sequence $\{(A_i, B_i, |\psi_i\rangle)\}_{i \in \mathbb{N}}$ where each element in the sequence belongs to $X_\Lambda^s \times X_\Lambda^s \times Y^s$. Note that since X_Λ^s and Y are two compact sets[8], $X_\Lambda^s \times X_\Lambda^s \times Y^s$ is also a compact set. Thus, the sequence of triples will have a limit point, call it $(A, B, |\psi\rangle)$, even if the sequence does not converge[9]. Defining $t' = \text{Prob}[B, |\psi\rangle] - \text{Prob}[A, |\psi\rangle]$ and noting that $\Lambda \mathbb{I} \geq B \geq A \geq 0$ we conclude that $t' \in K_\Lambda$. But we also know that the sequence $\{t'_i\}_{i \in \mathbb{N}}$ converges to t and therefore the limit point, $t' = t$ establishing that $t \in K_\Lambda$. □

Corollary 64. *For $\Lambda \in (0, \infty)$, $K_\Lambda = \{a \in V | \forall \Phi \in K_\Lambda^*, \Phi(a) \geq 0\}$. Further, $a \in K_\Lambda$ if and only if $\sum_x a(x) = 0$, $\sum_x x a(x) \geq 0$ and $\forall \lambda \in (-\infty, -\Lambda] \cup (0, \infty)$, $\sum_x \frac{\lambda x}{\lambda + x} a(x) \geq 0$.*

Proof. The first part is nearly trivial now. Since K_Λ is closed (see Lemma 63), $K_\Lambda^{**} = K_\Lambda$ (see Fact 61). We obtain the assertion by writing the dual cone (see Definition 55) of K_Λ^*.

The second part is also straight forward. Suppose we write the characterisation of operator monotone functions on $[0, \Lambda]$ (see Definition 57) so that any operator monotone function $f(x)$ can be written as $f(x) = \int d\omega(\lambda) f_\lambda(x)$ for some choice of $\omega(\lambda)$ where we have neglected for simplicity the c_0 and $c_1 x$ parts (they would not affect the argument). To check membership in K_Λ, as we just proved, it

[8] Compactness generalises the notion of a closed (i.e. a set that contains all its limit points) and bounded (i.e. a set where all its points lie at a fixed distance from each other) in Eucledian space.

[9] e.g. 0,1,0,2,0,3,0,4... has a limit point, zero; basically a subsequence with a limit. a sequence converges if all subsequences have the same limit point.

suffices to check that $\sum_{x\in\text{supp}(x)} f(x)a(x) \geq 0$ for all possible operator monotones $f(x)$. Using the characterisation of $f(x)$ the inequality translates to $\sum_{x\in\text{supp}(x)} d\omega(\lambda) f_\lambda(x)a(x) \geq 0$. Evidently, if this is to hold for all possible measures over λ, then it is necessary that $f_\lambda(x)a(x) \geq 0$ for each λ (else we could put all the weight on this λ and obtain a violation) and sufficient (because if each is non-negative, a convex sum would also be non-negative). This is how the extremal rays of the cone of operator monotone functions significantly simplify our task. □

A seemingly cumbersome (but useful) restatement of this result is the following.

Corollary 65 (EBM on $[0,\Lambda]$ is equivalent to Λ valid)**.** *A function $a : [0,\Lambda] \to \mathbb{R}$ with finite support on $[0,\Lambda]$ is EBM on $[0,\Lambda]$ if and only if a is Λ-valid, i.e., it satisfies $\sum_x a(x) = 0$, $\sum_x xa(x) \geq 0$ and $\forall\lambda \in (-\infty,-\Lambda]\cup(0,\infty)$, $\sum_x \frac{\lambda x}{\lambda+x}a(x) \geq 0$.*

Note that all the statements made here assume that the matrices used in EBM functions have a finite spectrum. Our EMA algorithm heavily relies on this part of the analysis which is due to Aharonov et al.

3.1.3 Strictly valid functions are EBM functions

To be able to simplify the conditions one needs to check, it is useful to remove the condition on the spectrum of the (positive semi-definite) matrices involved. This is evident from the range of λ one needs to use in the characterisation of operator monotone functions (compare Lemma 60 and Lemma 59).

It is easy to describe the interior of the dual of a cone. It is also possible to relate the interior with the closure of the cone, but in finite dimensions. This reasoning fails for infinite dimensions. They still serve as motivation for the definition of valid and strictly valid functions.

Fact 66. *[9] Let C be a convex set, then int(C) = int(cl(C)).*

Fact 67. *[9] Let C be a cone in the finite-dimensional vector space V, then int(C^*) = $\{\Phi \in V' | \forall a \in C\backslash\{0\}, \Phi(a) > 0\}$.*

It turns out that K is not closed[10], and recall that K^* is the set of operator monotone functions on $[0,\infty)$. Recall also that $K^{**} = \text{cl}(K)$. Using Fact 66 we conclude that $\text{int}(K^{**}) = \text{int}(K)$. If we pretend to be in finite dimensions, then restricting to the interior of K^{**} is easy as characterised in Fact 67. We could then simply consider points in the interior of K^* which in turn would guarantee membership in K. We are not in finite dimensions but the result continues to hold. To see this, we first define valid and strictly valid functions in the same vein.

Definition 68 (valid function)**.** A function $a : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with finite support is valid if for every operator monotone function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ we have $\sum_{x\in\text{supp}(h)} f(x)a(x) \geq 0$.

Definition 69 (strictly valid function)**.** A function $a : \mathbb{R}_{\geq 0} \to \mathbb{R}$ with finite support is strictly valid if for every non-constant operator monotone function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ we have $\sum_{x\in\text{supp}(a)} f(x)a(x) > 0$.

[10] One example is merge. Let $p_{g_1} + p_{g_2} = 1$, $x_{g_2} > x_{g_1} > 0$. Consider the sequence $t_1, t_2 \ldots t_k$ where $t_k := \left[\!\left[\langle x_g\rangle + \frac{1}{k}\right]\!\right] - p_{g_1}\left[\!\left[x_{g_1}\right]\!\right] - p_{g_2}\left[\!\left[x_{g_2}\right]\!\right]$. This sequence, in the limit $k \to \infty$ is just a merge. One can show that for any finite k, t_k can be shown to be EBM using matrices with a finite spectrum (this is because for a $2 \to 1$ transition, it suffices to restrict to 2×2 matrices (see Lemma 104) and then one can consider the most general unitary to reach the conclusion). However, as $k \to \infty$, the spectra of the matrices involved diverges. Thus, while the elements of the sequence $t_1, t_2 \ldots t_k \ldots$ are contained in K, its limit is not. This argument does not apply to K_Λ (confirming the fact that K_Λ is closed) because after some finite k, t_k ceases to be in K_Λ.

One can use the characterisation of operator monotone functions to explicitly characterise the set of valid and strictly valid functions (just as we did for Λ-valid functions; see Corollary 64).

Lemma 70. *Let $a : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a function with finite support such that $\sum_x a(x) = 0$. The function a is a strictly valid function if and only if for all $\lambda > 0$, $\sum_x \frac{-a(x)}{\lambda+x} > 0$ and $\sum_x x.a(x) > 0$.*

The function a is valid if and only if for all $\lambda > 0$, $\sum_x \frac{-a(x)}{\lambda+x} \geq 0$ and $\sum_x x.a(x) \geq 0$.

Proof. Using Lemma 59 and proceeding as in the proof of Corollary 64 we conclude that a is strictly valid if and only if

$$\sum_{x \in \text{supp}(a)} a(x) = 0,$$

for all $\lambda > 0$,

$$\sum_{x \in \text{supp}(a)} \frac{\lambda x}{\lambda + x} \cdot a(x) > 0, \tag{3.3}$$

and

$$\sum_{x \in \text{supp}(a)} x \cdot a(x) > 0. \tag{3.4}$$

To obtain the desired form[11], as we have done before, we observe that

$$\sum_x \frac{\lambda x}{\lambda + x} a(x) > 0 \iff \sum_x \left(1 + \frac{-\lambda}{\lambda + x}\right) a(x) > 0 \iff \sum_x \frac{-1}{\lambda + x} a(x) > 0.$$

The same argument goes through for the valid condition. □

The set of strictly valid functions can be shown to also be Λ valid for some finite Λ. This means that it would also be EBM on $[0, \Lambda]$ which in turn means it would be an EBM function. We hence have the following.

Lemma 71. *Any strictly valid function is an EBM function.*

Proof. Suppose a is some strictly valid function. We show that there is some $\Lambda > 0$ for which a is an EBM function on $[0, \Lambda]$ (and is hence an EBM function). To this end, recall that the characterisation of Corollary 65 is very similar to that of Lemma 70 (using the equivalence between the operator monotones $\frac{-1}{\lambda+x}$ and $\frac{\lambda x}{\lambda+x}$ in the characterisation). We effectively, only need to show that there is some $\Lambda > 0$ such that for all $\lambda < -\Lambda$, $\sum_x \frac{\lambda x}{\lambda+x} a(x) \geq 0$. To see this, consider the expression $\sum_x \frac{\lambda x}{\lambda+x} a(x)$ as a function of λ. We are given that for $\lambda > 0$, the expression is positive. As λ approaches $-\infty$, the expression approaches $\lim_{\lambda \to \infty} \sum_x \frac{\lambda x}{\lambda+x} a(x) = \sum_x x a(x) > 0$. Since the expression is continuous in λ, there must be a value Λ such that for all $\lambda < -\Lambda$, the expression is non-negative, i.e. $\sum_x \frac{\lambda x}{\lambda+x} a(x) \geq 0$. □

We conclude this subsection by similarly defining valid and strictly valid *transitions*.

Definition 72 (Valid and strictly valid line transitions)**.** Let $g, h : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be two functions with finite support. The transition $g \to h$ is valid (resp., strictly valid) if the function $h - g$ is valid (resp., strictly valid).

[11] It might seem that the last condition is redundant because in the limit of λ tending to infinity, we seem to recover Equation (3.4) from Equation (3.3). This is not correct because otherwise a two-point merge becomes a strictly valid function but there are no matrices with finite spectrum for which it is EBM; this is a contradiction because for each a that is strictly valid, there is a Λ for which $a \in K_\Lambda$ (see Lemma 71), i.e. there should exist matrices with their spectrum in $[0, \Lambda]$ which certify that a is EBM.

3.1.4 From point games with valid functions to point games with EBM functions

If we construct a point game with valid functions we can convert it into a game with EBM functions with an arbitrarily small overhead on the bias. The trick is to raise the coordinates of all the final points (ones with positive weight) a little at each step, to convert a valid function into a strictly valid function.

Theorem 73 (valid to EBM). *Given a point game with $2m$ valid functions and final point $[\beta, \alpha]$ and any $\epsilon > 0$, we can construct a point game with $2m$ EBM functions and final point $[\beta + \epsilon, \alpha + \epsilon]$.*

Proof. It suffices to show this same statement with strictly valid functions (which we demonstrate in Lemma 74) because strictly valid functions are EBM functions (see Lemma 71). □

Lemma 74. *Fix $\epsilon > 0$. Given a point game with $2m$ valid functions and final point $[\beta, \alpha]$ we can construct a point game with $2m$ strictly valid functions and final point $[\beta + \epsilon, \alpha + \epsilon]$.*

Proof. Suppose the $2m$ valid functions are $\{t_1, t_2 \dots t_{2m}\}$ and denote the corresponding strictly valid functions by $\{t'_1, t'_2, \dots t'_{2m}\}$. The idea is simply to raise each point by a factor of ϵ/m for both horizontal and vertical points so that the final point gets raised to $[\beta + \epsilon, \alpha + \epsilon]$ after the $2m$ functions.

For all $i \in \{1, 2 \dots 2m\}$ and for all $(x, y) \in \mathbb{R}_{\geq} \times \mathbb{R}_{\geq}$, consider the following functions

$$t'_i(x,y) = \begin{cases} t_i^+(x - i\epsilon/m, y - (i-1)\epsilon/m) - t_i^-(x - (i-1)\epsilon/m, y - (i-1)\epsilon/m) & \text{for } i \text{ odd} \\ t_i^+(x - (i-1)\epsilon/m, y - i\epsilon/m) - t_i^-(x - (i-1)\epsilon/m, y - (i-1)\epsilon/m) & \text{for } i \text{ even} \end{cases}$$

where t_i^+ and t_i^- are the positive and negative parts of t_i with disjoint support. Suppose i is odd (the other case follows analogously). We assert that $t'_i(x, y)$ is a strictly valid horizontal function. To see this, note that for all non-constant operator monotone functions and for all $y \in [0, \infty)$ we have

$$\begin{aligned}
\sum_{x \in \text{supp}(t_i'^+)} t_i'^+(x,y) f(x) &= \sum_{x \in \text{supp}(t'_i)} t_i^+(x - i\epsilon/m, y - (i-1)\epsilon/m) f(x) \\
&= \sum_{x \in \text{supp}(t_i^+)} t_i^+(x, y - (i-1)\epsilon/m) f(x + i\epsilon/m) && \text{from the definition of } t'_i \\
&\geq \sum_{x \in \text{supp}(t_i^-)} t_i^-(x, y - (i-1)\epsilon/m) f(x + i\epsilon/m) && \text{because } t_i \text{ is vertically valid} \\
&= \sum_{x \in \text{supp}(t_i^-)} t_i'^-(x + (i-1)\epsilon/m, y) f(x + i\epsilon/m) && \text{from the definition of } t'_i \\
&= \sum_{x \in \text{supp}(t_i'^-)} t_i'^-(x,y) f(x + \epsilon/m) && \text{again, from the definition of } t'_i \\
&> \sum_{x \in \text{supp}(t_i'^-)} t_i'^-(x,y) f(x) && \text{because } f(x + \epsilon/m) > f(x)
\end{aligned}$$

which establishes the assertion. It might help to note that if $x \in \text{supp}(t_i'^+)$ then $x - \frac{i\epsilon}{m} \in \text{supp}(t_i^+)$ and if $x \in \text{supp}(t_i'^-)$ then $x - (i-1)\frac{\epsilon}{m} \in \text{supp}(t_i^-)$. For i even, the analogous argument demonstrates that $t'_i(x, y)$ is a strictly valid vertical function.

□

II

Part

Primary Contributions

4

Chapter

TDPG-to-Explicit-protocol Framework (TEF) and bias 1/10

In this chapter, we give a framework for converting (time dependent) point games into explicit protocols granted an EBM like condition (see Definition 25) holds. We then use it to construct the unitaries which specify WCF protocols approaching bias $1/10$.

Notation. We use $\mathcal{N}(|\psi\rangle) := |\psi\rangle / \sqrt{\langle\psi|\psi\rangle}$.

§ 4.1 Motivation and Conventions

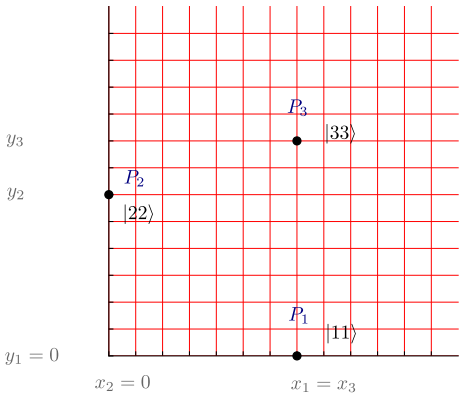
We wish to construct a protocol (see Definition 17) such that its dual matches (see Theorem 20) a given TDPG. The main difference in our construction, compared to the one used by Aharonov et al. and Mochon, is that the message register decouples after each round which is achieved by suitably placing the cheat-detection projectors. Consequently, the non-trivial constraint that the dual matrices must satisfy turns out to be similar to, but not exactly the same as, the EBM condition.

Keep Definition 24 in mind. Intuitively, the most natural way of constructing Zs and a $|\psi\rangle$ given an arbitrary frame (think of a TDPG as a sequence of frames) is to construct an entangled state that encodes the weight and define Zs to contain the coordinates corresponding to the weight. Let us make this idea more precise.

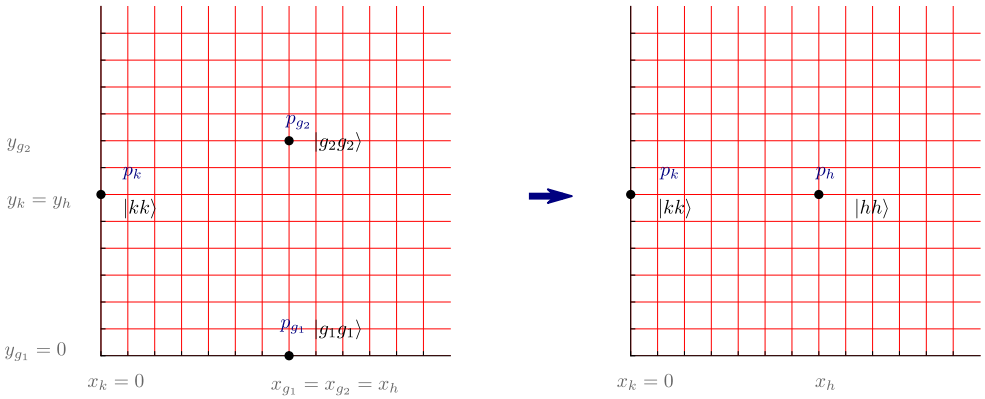
Definition (Canonical Form)**.** The tuple $(|\psi\rangle, Z^A, Z^B)$ is said to be in the Canonical Form with respect to a set of points in a frame of a TDPG if (see Figure 4.1a) $|\psi\rangle = \sum_i \sqrt{P_i}\,|ii\rangle_{AB} \otimes |\varphi\rangle_M$, $Z^A = (\sum x_i\,|i\rangle\langle i|_A)$ and $Z^B = (\sum y_i\,|i\rangle\langle i|_B)$ where $|\varphi\rangle_M$ represents the state of extra uncoupled registers which might be present.

It is easy to see that the 'label' $|ii\rangle$ correspond to a point with coordinates x_i, y_i and weight P_i in the frame (see Definition 24). It is tempting to imagine that we systematically construct, from each frame of a TDPG, a canonical form of $|\psi\rangle$ s and Zs. The unitaries can be deduced from the evolution of $|\psi\rangle$. This approach has two problems, (1) it does not manifestly mean that the unitaries would be decomposable into moves by Alice and Bob who communicate only through the message register and (2) the constraints imposed on consecutive Zs, of the form $Z_{n-1}\otimes\mathbb{I} \geq U_n^\dagger (Z_n \otimes \mathbb{I}) U_n$, are not satisfied in general. Our construction ensures these issues are dissolved.

The framework will output variables in the reverse time convention (as we did in Definition 28 and Proposition 29) indexed as, for example, $\left|\psi_{(i)}\right\rangle$, $Z_{(i)}$, $U_{(i)}$. The variables at the i^{th} step of the protocol (which follows the forward time convention) would be given by $|\psi_i\rangle = \left|\psi_{(N-i)}\right\rangle$, $Z_i = Z_{(N-i)}$ and $U_i = U_{(N-i)}^\dagger$. Note that the results so obtain extend naturally to the case where U_i may not be unitary and contains projections, e.g. $U_i E_i = E_{(N-i)} U_{(N-i)}^\dagger$.



(a) Frame of a TDPG



(b) The points which are unchanged from one frame to another are labelled by $\{k_i\}$. Among the points that change, the initial ones are labelled by $\{g_i\}$ and the final ones by $\{h_i\}$.

Figure 4.1: Illustrations for the Canonical Form

Basic Moves Work Out of the Box

Recall the three basic moves of a TDPG were given by

1. Raise: $p_1 [\![x, y]\!] \to p_1 [\![x', y]\!]$ s.t. $x' \geq x$ (see Example 31).

2. Merge: $p_1 [\![x_1, y]\!] + p_2 [\![x_2, y]\!] \to (p_1 + p_2) \left[\!\!\left[\frac{p_1 x_1 + p_2 x_2}{p_1 + p_2}, y \right]\!\!\right]$ (see Example 32).

3. Split: $(p_1 + p_2) \left[\!\!\left[\left(\frac{p_1 w_1 + p_2 w_2}{p_1 + p_2} \right)^{-1}, y \right]\!\!\right] \to p_1 [\![x_1, y]\!] + p_2 [\![x_2, y]\!]$ where $w_1 = 1/x_1$ and $w_2 = 1/x_2$ (see Example 33).

We construct the explicit Unitaries that implement these moves which in turn (when generalised to n points) are enough to construct the former best known protocol from its TDPG. Note, however, that these moves do not exhaust the set of moves and more advanced moves will be constructed to go beyond the $1/6$ limit.

§ 4.2 The Framework

Intuition

Imagine a canonical description is given. Let the labels on the points one wants to transform be indexed by $\{g_i\}$ and let us also assume that one wishes to apply an x-transition (i.e. Alice performs the non-trivial step). Let the labels of the points that one wishes to leave unchanged be given by $\{k_i\}$ (see Figure 4.1b). We can write the state as

$$|\psi_{(1)}\rangle = \left(\sum_i \sqrt{p_{g_i}}\,|g_i g_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}}\,|k_i k_i\rangle_{AB}\right) \otimes |m\rangle_M .$$

We want Bob to send his part of $|g_i\rangle$ states to Alice through the message register. One way is that he conditionally swaps to obtain the following

$$|\psi_{(2)}\rangle = \sum_i \sqrt{p_{g_i}}\,|g_i g_i\rangle_{AM} \otimes |m\rangle_B + \sum_i \sqrt{p_{k_i}}\,|k_i k_i\rangle_{AB} \otimes |m\rangle_M .$$

This should at most force all the points to align along the y-axis but no non-trivial constraint should arise (speaking with hindsight). Let $\{h_i\}$ be the labels of the new points after the transformation. We assume that h_i, g_i and k_i index orthonormal vectors. Alice can update the probabilities and labels by locally performing a unitary to obtain

$$|\psi_{(3)}\rangle = \sum_i \sqrt{p_{h_i}}\,|h_i h_i\rangle_{AM} \otimes |m\rangle_B + \sum_i \sqrt{p_{k_i}}\,|k_i k_i\rangle_{AB} \otimes |m\rangle_M .$$

It is precisely this step which yields the non-trivial constraint. Bob must now accept this by 'unswapping' to get

$$|\psi_{(4)}\rangle = \left(\sum_i \sqrt{p_{h_i}}\,|h_i h_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}}\,|k_i k_i\rangle_{AB}\right) \otimes |m\rangle_M$$

which leaves Bob's Z in essentially the standard form (we will see). Remember that in the actual protocol the sequence will get reversed as described above (see Remark 19, Remark 21 and Equation (A.2)).

Note that we add a few extra frames to the final TDPG to go from a given frame to the next of the original TDPG. This is irrelevant, when resource usage is not of interest, as the bias stays the same but we mention it to avoid confusion.

Formal Description and Proofs

1. **First frame.**

$$\begin{aligned}
|\psi_{(1)}\rangle &= \left(\sum_i \sqrt{p_{g_i}}\,|g_i g_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}}\,|k_i k_i\rangle_{AB}\right) \otimes |m\rangle_M \\
Z^A_{(1)} &= \sum_i x_{g_i}\,|g_i\rangle\langle g_i|_A + \sum_i x_{k_i}\,|k_i\rangle\langle k_i|_A \\
Z^B_{(1)} &= \sum_i y_{g_i}\,|g_i\rangle\langle g_i|_B + \sum_i y_{k_i}\,|k_i\rangle\langle k_i|_B .
\end{aligned}$$

Proof. Follows from the assumption of starting with a Canonical Form. □

2. **Bob sends to Alice.** With $y \geq \max\{y_{g_i}\}$ the following is a valid choice

$$\begin{aligned}
\left|\psi_{(2)}\right\rangle &= \sum_i \sqrt{p_{g_i}} \left|g_i g_i\right\rangle_{AM} \otimes \left|m\right\rangle_B + \sum_i \sqrt{p_{k_i}} \left|k_i k_i\right\rangle_{AB} \otimes \left|m\right\rangle_M \\
U^{(1)} &= U_{BM}^{\mathrm{SWP}\{\vec{g},m\}} \\
Z_{(2)}^A &= Z_{(1)}^A \\
Z_{(2)}^B &= y\mathbb{I}_B^{\{\vec{g},m\}} + \sum_i y_{k_i} \left|k_i\right\rangle \left\langle k_i\right|_B .
\end{aligned}$$

Proof. We have to prove: (1) $\left|\psi_{(2)}\right\rangle = U^{(1)} \left|\psi_{(1)}\right\rangle$ and (2) $U^{(1)\dagger} \left(Z_{(2)}^B \otimes \mathbb{I}_M\right) U^{(1)} \geq \left(Z_{(1)}^B \otimes \mathbb{I}_M\right)$.
(1) It follows trivially from the defining action of $U^{(1)}$.
(2) For convenience, let momentarily $U = U^{(1)}$ and note that $U^\dagger = U$ so that we can write

$$\begin{aligned}
&U \left(Z_{(2)}^B \otimes \mathbb{I}_M\right) U \\
&= y \left(U \left(\mathbb{I}_B^{\{\vec{g},m\}} \otimes \mathbb{I}_M^{\{\vec{g},m\}}\right) U + U \underbrace{\left(\mathbb{I}_B^{\{\vec{g},m\}} \otimes \mathbb{I}_M^{\{\vec{k},\vec{h}\}}\right)}_{\text{outside } U\text{'s action space}} U \right) + U \underbrace{\left(\sum y_{k_i} \left|k_i\right\rangle \left\langle k_i\right| \otimes \mathbb{I}\right)}_{\text{outside } U\text{'s action space}} U \\
&= Z_{(2)} \otimes \mathbb{I}_M \geq Z_{(1)} \otimes \mathbb{I}_M
\end{aligned}$$

so long as $y \geq y_{g_i}$ which is guaranteed by the choice of y. □

3. **Alice's non-trivial step.** We claim that the following is a valid choice,

$$\begin{aligned}
\left|\psi_{(3)}\right\rangle &= \sum_i \sqrt{p_{h_i}} \left|h_i h_i\right\rangle_{AM} \otimes \left|m\right\rangle_B + \sum_i \sqrt{p_{k_i}} \left|k_i k_i\right\rangle_{AB} \otimes \left|m\right\rangle_M \\
E^{(2)}U^{(2)} &= E^{(2)} \left(\left|w\right\rangle \left\langle v\right| + \text{other terms acting on span}\{\left|h_i h_i\right\rangle, \left|g_i g_i\right\rangle\}\right)_{AM} \\
Z_{(3)}^A &= \sum_i x_{h_i} \left|h_i\right\rangle \left\langle h_i\right| + \sum_i x_{k_i} \left|k_i\right\rangle \left\langle k_i\right| \\
Z_{(3)}^B &= Z_{(2)}^B
\end{aligned}$$

where

$$\left|v\right\rangle = \frac{\sum_i \sqrt{p_{g_i}} \left|g_i g_i\right\rangle}{\sqrt{\sum_i p_{g_i}}},\ \left|w\right\rangle = \frac{\sum_i \sqrt{p_{h_i}} \left|h_i h_i\right\rangle}{\sqrt{\sum_i p_{h_i}}},\ E^{(2)} = \left(\sum \left|h_i\right\rangle \left\langle h_i\right|_A + \sum \left|k_i\right\rangle \left\langle k_i\right|_A\right) \otimes \mathbb{I}_M$$

subject to the condition

$$\sum x_{h_i} \left|h_i h_i\right\rangle \left\langle h_i h_i\right| \geq \sum x_{g_i} E^{(2)} U^{(2)} \left|g_i g_i\right\rangle \left\langle g_i g_i\right| U^{(2)\dagger} E^{(2)} \tag{4.1}$$

and of course the conservation of probability, viz. $\sum p_{g_i} = \sum p_{h_i}$.

Proof. We must show that (1) $E^{(2)} \left|\psi_{(3)}\right\rangle = U^{(2)} \left|\psi_{(2)}\right\rangle$ and (2)

$$Z_{(3)}^A \otimes \mathbb{I}_M \geq E^{(2)} U^{(2)} \left(Z_{(2)}^A \otimes \mathbb{I}_M\right) U^{(2)\dagger} E^{(2)}.$$

(1) Observing $E^{(2)} \left|\psi_{(3)}\right\rangle = \left|\psi_{(3)}\right\rangle$ the statement holds almost trivially by construction of $U^{(2)}$.
(2) Consider the space $\mathcal{H} = \text{span}\{\left|g_1 g_1\right\rangle, \left|g_2 g_2\right\rangle \ldots, \left|h_1 h_1\right\rangle, \left|h_2, h_2\right\rangle \ldots\}$. We separate all expressions (they are nearly diagonal) into the $\mathcal{H}$ space (which gets non-diagonal) and the rest. We start with the RHS, excluding the Us,

$$Z_{(2)}^A \otimes \mathbb{I}_M = \underbrace{\sum x_{g_i} \left|g_i g_i\right\rangle \left\langle g_i g_i\right|}_{\mathrm{I}} + \sum x_{g_i} \left|g_i\right\rangle \left\langle g_i\right| \otimes (\mathbb{I} - \left|g_i\right\rangle \left\langle g_i\right|) + \sum x_{k_i} \left|k_i\right\rangle \left\langle k_i\right| \otimes \mathbb{I},$$

where only term I is in the operator space spanned by $\mathcal{H}$. Note that all the terms are still diagonal. Next consider the LHS,

$$Z^A_{(3)} \otimes \mathbb{I}_M = \underbrace{\sum x_{h_i} |h_i h_i\rangle \langle h_i h_i|}_{\text{I}} + \sum x_{h_i} |h_i\rangle \langle h_i| \otimes (\mathbb{I} - |h_i\rangle \langle h_i|) + \sum x_{k_i} |k_i\rangle \langle k_i| \otimes \mathbb{I},$$

which also has only term I in the $\mathcal{H}$ operator space. Consequently, only on these will U have a non-trivial action. Let us first evaluate the non-$\mathcal{H}$ part where we only need to apply the projector. The result after separating equations where possible is

$$\sum x_{h_i} |h_i\rangle \langle h_i| \otimes (\mathbb{I} - |h_i\rangle \langle h_i|) \geq 0$$
$$\sum (x_{k_i} - x_{k_i}) |k_i\rangle \langle k_i| \otimes \mathbb{I} \geq 0$$

which essentially only implies

$$x_{h_i} \geq 0.$$

Finally the non-trivial part yields

$$\sum x_{h_i} |h_i h_i\rangle \langle h_i h_i| \geq \sum x_{g_i} EU |g_i g_i\rangle \langle g_i g_i| U^\dagger E$$

which completes the proof. □

4. **Bob accepts Alice's change.** The following is valid.

$$\begin{aligned}
|\psi_{(4)}\rangle &= \left(\sum_i \sqrt{p_{h_i}} |h_i h_i\rangle_{AB} + \sum_i \sqrt{p_{k_i}} |k_i k_i\rangle_{AB} \right) \otimes |m\rangle_M \\
E^{(3)} U^{(3)} &= E^{(3)} U^{\text{SWP}\{\vec{h},m\}}_{BM} \\
Z^A_{(4)} &= Z^A_{(3)} \\
Z^B_{(4)} &= y \sum_i |h_i\rangle \langle h_i| + \sum_i y_{k_i} |k_i\rangle \langle k_i|_B
\end{aligned}$$

where $E^{(3)} = (\sum |h_i\rangle \langle h_i| + \sum |k_i\rangle \langle k_i|)_B \otimes \mathbb{I}_M$.

Proof. We have to prove: (1) $E^{(3)} |\psi_{(4)}\rangle = U^{(3)} |\psi_{(3)}\rangle$ and (2)

$$Z^B_{(4)} \otimes \mathbb{I}_M \geq E^{(3)} U^{(3)} \left(Z^B_{(3)} \otimes \mathbb{I}_M \right) U^{(3)\dagger} E^{(3)}.$$

(1) This can be proven again, by a direct application of $U^\dagger E$ on $|\psi_{(4)}\rangle$ (where E is defined to be $E^{(3)}$ and U to be $U^{(3)}$ for the proof).
(2) Note that

$$\begin{aligned}
EU \left(\mathbb{I}_B^{\{\vec{g},m\}} \otimes \mathbb{I}_M^{\{\vec{h},\vec{g},\vec{k},m\}} \right) U^\dagger E &= EU \left(\mathbb{I}_B^{\{m\}} \otimes \mathbb{I}_M^{\{\vec{h},\vec{g},\vec{k},m\}} \right) U^\dagger E + E \left(\mathbb{I}_B^{\{\vec{g}\}} \otimes \mathbb{I}_M^{\{\vec{h},\vec{g},\vec{k},m\}} \right) E \\
&= EU \left(\mathbb{I}_B^{\{m\}} \otimes \mathbb{I}_M^{\{\vec{h},m\}} \right) U^\dagger E \\
&= \sum |h_i\rangle \langle h_i| \otimes \mathbb{I}_M^{\{m\}}.
\end{aligned}$$

Since the other term in $Z^B_3 \otimes \mathbb{I}$ is anyway in the non-action space of U it follows that

$$EU(Z^B_3 \otimes \mathbb{I}) U^\dagger E = y \sum |h_i\rangle \langle h_i| \otimes \mathbb{I}_M^{\{m\}} + \sum y_{k_i} |k_i\rangle \langle k_i| \otimes \mathbb{I}_M.$$

It only remains to show that $Z^B_{(4)} \otimes \mathbb{I}_M \geq E^{(3)} U^{(3)} \left(Z^B_{(3)} \otimes \mathbb{I}_M \right) U^{(3)\dagger} E^{(3)}$ which it obviously is because $y \sum |h_i\rangle \langle h_i| \otimes \mathbb{I}_M \geq y \sum |h_i\rangle \langle h_i| \otimes \mathbb{I}_M^{\{m\}}$ and the y_{k_i} term is common. □

We can summarise the condition of interest as follows, the proof of which is a trivial consequence of the aforesaid.

Theorem 75. *For an x-transition (where Alice performs the non-trivial step)*

$$\sum_{i=1}^{n_k} p_{k_i} [\![x_{k_i}]\!] + \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!] \to \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!] + \sum_{i=1}^{n_k} p_{k_i} [\![x_{k_i}]\!]$$

to be implementable under the TDPG-to-Explicit-protocol Framework (TEF) it suffices find a $U^{(2)}$ that satisfies the inequality

$$\sum_{i=1}^{n_h} x_{h_i} |h_i h_i\rangle \langle h_i h_i|_{AM} \geq \sum_{i=1}^{n_g} x_{g_i} E_h^{(2)} U^{(2)} |g_i g_i\rangle \langle g_i g_i|_{AM} U^{(2)\dagger} E_h^{(2)} \tag{4.2}$$

and the honest action constraint

$$U^{(2)} |v\rangle = |w\rangle$$

where $|h_i\rangle$ and $|g_i\rangle$ are orthonormal basis vectors,

$$|v\rangle = \mathcal{N}\left(\sum \sqrt{p_{g_i}} |g_i g_i\rangle_{AM}\right)$$

and

$$|w\rangle = \mathcal{N}\left(\sum \sqrt{p_{h_i}} |h_i h_i\rangle_{AM}\right)$$

for $\mathcal{N}(|\psi\rangle) = |\psi\rangle / \sqrt{\langle\psi|\psi\rangle}$, $E_h = (\sum_{i=1}^{n_h} |h_i\rangle \langle h_i|_A + \sum |k_i\rangle \langle k_i|_A) \otimes \mathbb{I}_M$ with $U^{(2)}$'s non-trivial action restricted to span $\{\{|g_i g_i\rangle_{AM}\}, \{|h_i h_i\rangle_{AM}\}\}$ (note $|k_i\rangle$ corresponds to the points that are left unchanged in the transition).

Theorem 75 also leads to Theorem 30. To this end, we need to show that the EBM condition implies the inequality which appears in Theorem 75 can be satisfied. There are two difficulties, the first is that in Equation (4.2), there is a projector and the second is that the matrices have a certain dimension while for EBM there was neither a projector, nor a constraint on the dimension. We address them in Chapter 6. It turns out that the projector can be seen to be the limiting case of one of the matrices having diverging eigenvalues (see Subsection 6.1.1). It also turns out that restricting to real unitaries (orthogonal matrices) does not lead to a loss of generality (see Corollary 102). Finally, and this Mochon had already proven, it suffices to have matrices of size $n_h + n_g - 1$ (see Lemma 104). Together, these establish Theorem 30.

In fact, the set of functions which satisfy the TEF constraints (for some $U^{(2)}$) as described above in Theorem 75 (or equivalently in Theorem 10) is the closure of the set of EBM functions, which in turn, is the set of valid functions (see Section B.1). Thus, using TEF, we can directly associate valid games with WCF protocols, granted of course, the non-trivial unitary, $U^{(2)}$ can be found. This, as a by-product, lets us skip the notion of strictly valid functions.

4.2.1 Important Special Case: The Blinkered Unitary

So far we have not specified the non-trivial $U^{(2)}$ (which we call U from now) beyond requiring it to have a certain action on the honest state. We now define an important class of U, we call the Blinkered Unitary, as

$$U = |w\rangle \langle v| + |v\rangle \langle w| + \sum_i |v_i\rangle \langle v_i| + \sum_i |w_i\rangle \langle w_i| + \mathbb{I}^{\text{outside } \mathcal{H}}$$

and can even drop the last term as we are restricting our analysis to the $\mathcal{H}$ operator space, where $|v\rangle, \{|v_i\rangle\}$ form a complete orthonormal basis and so do $|w\rangle, \{|w_i\rangle\}$ wrt span$\{|g_i g_i\rangle\}$ and span$\{|v_i v_i\rangle\}$ respectively. The blinkered unitary can be used to implement the two non-trivial operations of the set of basic moves.

- Merge: $g_1, g_2 \to h_1$
  We can construct from the very definitions

$$|v\rangle = \frac{\sqrt{p_{g_1}}|g_1 g_1\rangle + \sqrt{p_{g_2}}|g_2 g_2\rangle}{N},\ |v_1\rangle = \frac{\sqrt{p_{g_2}}|g_1 g_1\rangle - \sqrt{p_{g_1}}|g_2 g_2\rangle}{N},\ |w\rangle = |h_1 h_1\rangle$$

  with $N = \sqrt{p_{g_1} + p_{g_2}}$ and even

$$U = |w\rangle\langle v| + |v\rangle\langle w| + |v_1\rangle\langle v_1|\ (= U^\dagger).$$

  We would need

$$EU|g_1 g_1\rangle = \frac{\sqrt{p_{g_1}}|w\rangle}{N},\ EU|g_2 g_2\rangle = \frac{\sqrt{p_{g_2}}|w\rangle}{N}$$

  because the constraint was (substituting for m and n)

$$x_h|h_1 h_1\rangle\langle h_1 h_1| \geq \sum x_{g_i} EU|g_i g_i\rangle\langle g_i g_i| U^\dagger E$$

  which becomes

$$x_h \geq \frac{p_{g_1} x_{g_1} + p_{g_2} x_{g_2}}{N^2}.$$

  This is precisely the merge condition (see Example 32). This can be readily generalised to an $m \to 1$ point merge condition by simply constructing the appropriate vectors (which we leave for the appendix).

- Split: $g_1 \to h_1, h_2$

$$|v\rangle = |g_1 g_1\rangle,\ |w\rangle = \frac{\sqrt{p_{h_1}}|h_1 h_1\rangle + \sqrt{p_{h_2}}|h_2 h_2\rangle}{N},\ |w_1\rangle = \frac{\sqrt{p_{h_2}}|h_1 h_1\rangle - \sqrt{p_{h_1}}|h_2 h_2\rangle}{N}$$

  with $N = \sqrt{p_{h_1} + p_{h_2}}$ and

$$U = |v\rangle\langle w| + |w\rangle\langle v| + |w_1\rangle\langle w_1| = U^\dagger.$$

  We evaluate $EU|g_1 g_1\rangle = |w\rangle$ which upon being plugged into the constraint yields

$$x_{h_1}|h_1 h_1\rangle\langle h_1 h_1| + x_{h_2}|h_2 h_2\rangle\langle h_2 h_2| - x_{g_1}|w\rangle\langle w| \geq 0.$$

  This yields the matrix equation

$$\begin{bmatrix} x_{h_1} & \\ & x_{h_2} \end{bmatrix} - \frac{x_{g_1}}{N^2}\begin{bmatrix} p_{h_1} & \sqrt{p_{h_1} p_{h_2}} \\ \sqrt{p_{h_1} p_{h_2}} & p_{h_2} \end{bmatrix} \geq 0$$

$$\mathbb{I} \geq \frac{x_{g_1}}{N^2}\begin{bmatrix} \frac{p_{h_1}}{x_{h_1}} & \sqrt{\frac{p_{h_1}}{x_{h_1}}\frac{p_{h_2}}{x_{h_2}}} \\ \sqrt{\frac{p_{h_1}}{x_{h_1}}\frac{p_{h_2}}{x_{h_2}}} & \frac{p_{h_2}}{x_{h_2}} \end{bmatrix}$$

$$\frac{x_{g_1}}{N^2}\left(\frac{p_{h_1}}{x_{h_1}} + \frac{p_{h_2}}{x_{h_2}}\right) \leq 1$$

  where in the second step we used the fact that $F - M \geq 0$ implies $\mathbb{I} - \sqrt{F}^{-1} M \sqrt{F}^{-1} \geq 0$ (if $F > 0$) and the last step is obtained by writing the matrix in the previous step as $|\psi\rangle\langle\psi|$ followed by demanding $1 \geq \langle\psi|\psi\rangle$.
  The last statement is the same constraint for a split (see Example 33). This also readily generalises to the case of $1 \to N$ splits which again we defer to the appendix.

  It would not be surprising to learn/prove that the class of unitaries with these properties is much more general than the Blinkered Unitaries.

- General $m \to n$: $g_1, g_2 \ldots g_m \to h_1, h_2 \ldots h_n$
  It is not too hard to show that in general one obtains the constraint

  $$\frac{1}{\langle x_g \rangle} \geq \left\langle \frac{1}{x_h} \right\rangle$$

  or more explicitly,

  $$\frac{1}{\sum_{i=1}^{m} p_{g_i} x_{g_i}} \geq \sum_{i=1}^{n} p_{h_i} . \frac{1}{x_{h_i}},$$

  using the appropriate blinkered unitary (which also we show in the appendix).

This class of unitary is enough to convert the $1/6$ game into an explicit protocol. However, for games given by Mochon that go beyond $1/6$ this class falls short. One way of seeing this is that the general $m \to n$ blinkered transition effectively behaves like an $m \to 1$ merge followed by a $1 \to n$ split, which are a set of moves that are insufficient to break the $1/6$ limit (at least using Mochon's games).

§ 4.3 Games and Protocols

We now describe two games, the bias $1/6$ game and the bias $1/10$ game, from the family of games constructed by Mochon to show that arbitrarily small bias is achievable (see Section 2.5). Recall that Mochon parametrised his games by k which determined the number of points involved in the non-trivial step. The bias he obtained was $\epsilon = 1/(4k+2)$. We consider games with $k = 1$ and $k = 2$, yielding the aforementioned bias.

4.3.1 Mochon's Approach

4.3.1.1 Assignments

Recall that a function

$$\sum_{z \in \{x_1, x_2, \ldots x_n\}} p(z) \, [\![z]\!]$$

is valid if

$$\sum_{z \in \{x_1, x_2, \ldots x_n\}} \left(\frac{-1}{\lambda + z} \right) p(z) \geq 0, \quad \sum_{z \in \{x_1, x_2, \ldots x_n\}} z p(z) \geq 0, \quad \sum_{z \in \{x_1 \ldots x_n\}} p(z) = 0$$

for all $\lambda > 0$ where $x_i \geq 0$. As we said before stating Lemma 45, checking if a generic assignment for p satisfies these infinite constraints in not always easy. Mochon had used a constructive approach here and we build on to it. Let us restate these results with some precision (proven in the appendix) where, as above, n distinct numbers are assumed to be represented by x_is and each $x_i \geq 0$.

Lemma 76 (Mochon's Denominator). $\sum_{i=1}^{n} \frac{1}{\prod_{j \neq i}(x_j - x_i)} = 0$ *for* $n \geq 2$ *where* $x_i \in \mathbb{R}$ *are distinct.*

Lemma 77 (Mochon's f-assignment Lemma). $\sum_{i=1}^{n} \frac{f(x_i)}{\prod_{j \neq i}(x_j - x_i)} = 0$ *where* $f(x_i)$ *is a polynomial of order* $k \leq n-2$ *where* $x_i \in \mathbb{R}$ *are distinct.*

Definition (Mochon's TIPG assignment)**.** Given a set of n points $0 \leq x_1 < x_2 \cdots < x_n$, a polynomial $f(x)$ with order k at most $n-2$ and $f(-\lambda) \geq 0$ for all $\lambda \geq 0$, the probability weights for a TIPG assignment is $p(x_i) = -\frac{f(x_i)}{\prod_{j \neq i}(x_j - x_i)}$.

Mochon was able to show that 'Mochon's TIPG assignment' makes for a valid function (in the TIPG formalism).

Proposition. *Suppose Mochon's TIPG assignment for $x_1 < x_2 \cdots < x_n$ is given by $p(x_i)$. Then, $\sum_{i=1}^{n} p(x_i) [\![x_i, y]\!]$ is a valid function for every $y \geq 0$.*

As we saw in the proof of Lemma 46, the power of this construction lies in the fact that we can easily construct polynomials that have roots at arbitrary locations. This allowed us to create interesting repeating structures called ladders (due to Mochon) which were terminated using these polynomials to obtain a game with a finite set of points. These ladders played a pivotal role in achieving smaller biases and the ability to obtain finite ladders is essential for being able to obtain a physical process that would yield the said bias.

We now build a little on Mochon's notation and results.

Definition (Mochon's TDPG assignment)**.** Given Mochon's TIPG assignment, let I be the set of indices for which $p(x_i) < 0$ and K be the remaining indices with respect to $\{1, 2, \ldots n\}$. The TDPG assignment (in accordance with the notation used in TEF) is given as

$$
\begin{aligned}
(x_{g_1}, x_{g_2} \ldots) &:= (x_i)_{i \in I} \\
(p_{g_1}, p_{g_2} \ldots) &:= (-p(x_i))_{i \in I} \\
(x_{h_1}, x_{h_2} \ldots) &:= (x_k)_{k \in K} \\
(p_{h_1}, p_{h_2} \ldots) &:= (p(x_k))_{k \in K}.
\end{aligned}
$$

With these in place we make some observations about how initial and final averages behave under such an assignment.

Proposition. *$N_h^2 = N_g^2$ where $N_g^2 = \sum p_{g_i}$ and $N_h^2 = \sum p_{h_i}$ for Mochon's TDPG assignment.*

Proof. We have to show that $N_h^2 - N_g^2 = \sum p_{h_i} - \sum p_{g_i} = 0$ which is the same as showing $\sum_{i=1}^{n} p(x_i) = 0$ which holds because we just saw that $\sum_{i=1}^{n} f(x_i) / \prod_{j \neq i}(x_j - x_i) = 0$ (Mochon's f-assignment Lemma). □

Proposition. *$\langle x_h \rangle - \langle x_g \rangle = 0$ for Mochon's TDPG assignment with $k \leq n-3$ where $\langle x_h \rangle = \frac{1}{N_h^2} \sum p_{h_i} x_{h_i}$ and $\langle x_g \rangle = \frac{1}{N_g^2} \sum p_{g_i} x_{g_i}$.*

Proof. This is a direct consequence of Mochon's f-assignment lemma. Let h be the $n-3$ order polynomial defined by Mochon's TDPG assignment so that $\langle x_h \rangle - \langle x_g \rangle \propto \sum p_{h_i} x_i - \sum p_{g_i} x_{g_i} = \sum_{i=1}^{n} p(x_i) x_i = \sum_{i=1}^{n} \frac{h(x_i) x_i}{\prod_{j \neq i}(x_j - x_i)} = \sum_{i=1}^{n} \frac{f(x_i)}{\prod_{j \neq i}(x_j - x_i)} = 0$ because f is an $n-2$ order polynomial. □

Lemma 78. *We have $\sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j \neq i}(x_j - x_i)} = (-1)^{n-1}$ for $n \geq 2$ (proof in Section B.3 of the Appendix).*

Proposition 79. *$\langle x_h \rangle - \langle x_g \rangle = \frac{1}{N_h^2} = \frac{1}{N_g^2}$ for a Mochon's TDPG assignment with $k = n-2$ and coefficient of x^{n-2} being ± 1 in $f(x)$. As above, here $\langle x_h \rangle = \frac{1}{N_h^2} \sum p_{h_i} x_{h_i}$ and $\langle x_g \rangle = \frac{1}{N_g^2} \sum p_{g_i} x_{g_i}$.*

We will see that typically $N_h = N_g$ are quite large and the average only slightly increase, if at all. We are now in a position to discuss Mochon's games.

4.3.1.2 Typical Game Structure

We assume an equally spaced n-point lattice given by $x_j = x_0 + j\delta x$ where $\delta x = \delta y$ is small and x_0 would essentially give a bound on P_B^* which will be determined by following the constraints; similarly $y_j = y_0 + j\delta y$ and we also define $\Gamma_{k+1} = y_{n-k} = x_{n-k}$. Let $P(x_j)$ be the probability weight associated with the point $[x_j, 0]$ s.t.

$$\sum_{j=1}^{n} P(x_j) = \frac{1}{2}, \sum_{j=1}^{n} \frac{P(x_j)}{x_j} = \frac{1}{2}.$$

Similarly with the point $(0, y_j)$ we associate $P(y_j)$ where $y_j = x_j$ as we also assume that $x_0 = y_0$. These choices explicitly impose symmetry between Alice and Bob which in turn entails that we have to do only half the analysis.

4.3.2 Bias $1/6$

4.3.2.1 Game

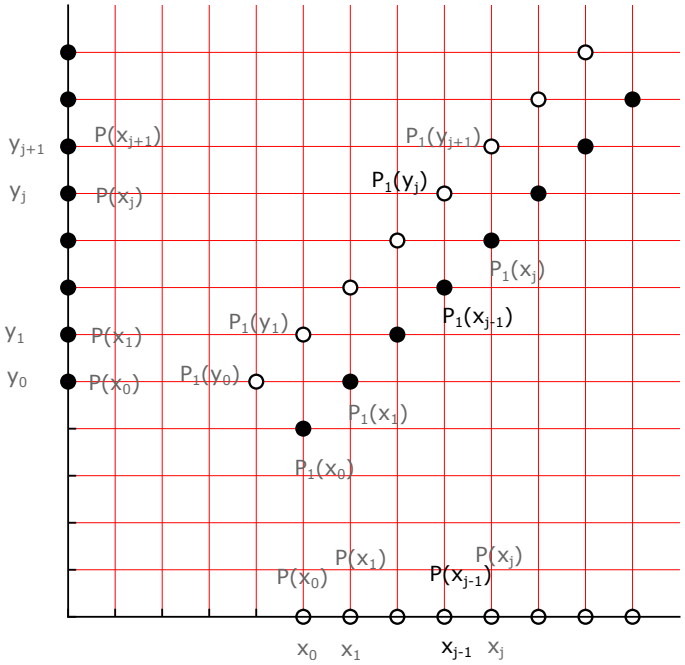


Figure 4.2: Building a TDPG/TIPG using merge moves

With reference to Figure 4.2 we need to satisfy $P(x_{j-1}) + P_1(y_j) = P_1(x_{j-1})$ which is probability conservation and $P_1(y_j)y_j \leq P_1(x_{j-1})y_{j-2}$ which is the merge condition. Both of these are automatically satisfied if we make a Mochon's denominator based assignment as follows

$$0 \leftrightarrow x_{g_1}$$
$$y_j \leftrightarrow x_{g_2}$$
$$y_{j-2} \leftrightarrow x_{h_1}$$

$$P(x_{j-1}) \leftrightarrow p_{g_1} = \frac{c(x_{j-1})}{y_j y_{j-2}}$$
$$P_1(y_j) \leftrightarrow p_{g_2} = \frac{c(x_{j-1})}{(y_j - y_{j-2})(y_j)} = \frac{c(x_{j-1})}{2y_j \delta y}$$
$$P_1(x_{j-1}) \leftrightarrow p_{h_1} = \frac{c(x_{j-1})}{(y_j - y_{j-2})(y_{j-2})} = \frac{c(x_{j-1})}{2y_{j-2} \delta y}$$

where the function $c(x_{j-1})$ must be chosen so that $P_1(y_j) = P_1(x_j)$ which entails

$$\frac{c(x_{j-1})}{\cancel{2}y_j \cancel{\delta y}} = \frac{c(x_j)}{\cancel{2}y_{j-1} \cancel{\delta y}}$$

and that in turn is solved by $c(x_j) = \frac{c_0 \delta x}{x_j}$ where we used $x_j = y_j$, $\delta x = \delta y$ (and added a δx as it helps approximating $\sum P(x_j)$ by an integral). Plugging this back we have

$$P_1(x_j) = \frac{c_0}{2x_j x_{j-1}}, \; P(x_j) = \frac{c_0 \delta x}{x_{j-1} x_j x_{j+1}}.$$

Since they involve a sum we do this in the limit $\delta x \to 0$ and $\Gamma \to \infty$ to avoid dealing with summing a series.

$$\sum_{j=0}^{n} P(x_j) = \frac{1}{2} \to c_0 \int_{x_0}^{\Gamma} \frac{dx}{x^3} = \frac{c_0}{(-2)} \left[\frac{1}{\Gamma^2} - \frac{1}{x_0^2} \right] = \frac{1}{2}$$

which entails $c_0 = x_0^2$. The next condition yields x_0

$$\sum_{j=0}^{n} \frac{P(x_j)}{x_j} = \frac{1}{2} \to x_0^2 \int_{x_0}^{\Gamma} \frac{dx}{x^4} = \frac{x_0^2}{(-3)} \left[\frac{1}{\Gamma^3} - \frac{1}{x_0^3} \right] = \frac{1}{3x_0} = \frac{1}{2}$$

which means

$$x_0 = \frac{2}{3} \implies \epsilon = \frac{1}{6}.$$

Of course a more careful analysis must be done to show these things exactly. Aside from the integration step one must also set $c_0(x) = (\Gamma_{n+1} - x)$ in order to terminate the ladder which turns the terminating step on the ladder into a raise. At the moment, however, we satisfy ourselves with this and move on to the more interesting $1/10$ game. We already saw how to deal with these issues in Chapter 2 supplemented by Appendix A.

4.3.2.2 Protocol

Although we could only claim that one can construct the protocol once the unitaries are known, the basic idea is that one starts with a split, then a raise by Alice and Bob, followed by a merge by Bob, then a merge by Alice and so on until only two points remain.[1] Bob can also start as the description is symmetric. These two can then be raised to the same location and merged. The coordinates of these points tend to $[\frac{2}{3}, \frac{2}{3}]$ as calculated above. The only creative part left would be the choice of labels that make the description neater from the point of view of the explicit protocol.

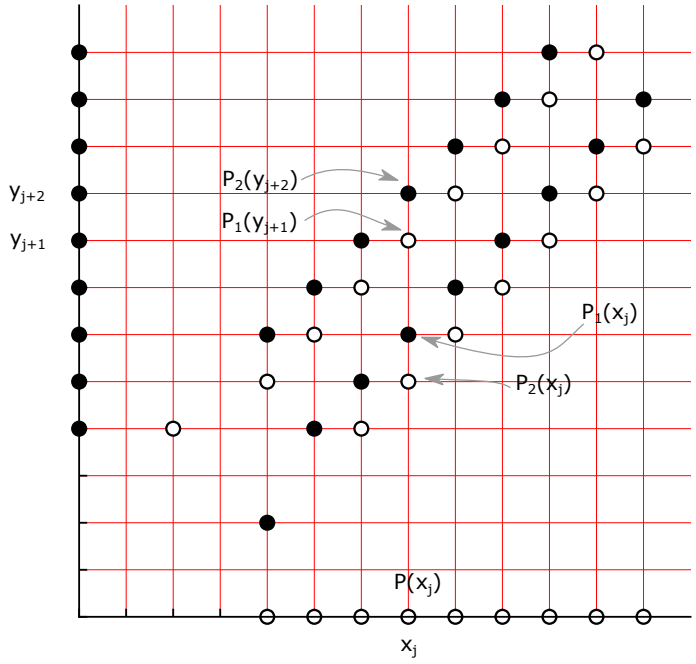


Figure 4.3: $1/10$ game: The $3 \to 2$ move based TIPG for bias $1/10$

4.3.3 Bias $1/10$ Game

With respect to Figure 4.3 we use Mochon's assignment with $f(y_i) = (y_{-2} - y_i)(\Gamma_1 - y_i)(\Gamma_2 - y_i)$ as

$$\frac{f(y_j)c(x_j)}{\prod_{k\neq j}(y_k - y_j)}.$$

Following the scheme as described above the probabilities become

$$\begin{aligned}
P_2(y_{j+2}) &= \frac{-f(y_{j+2})c(x_j)}{4.3(\delta y)^2 y_{j+2}} \\
P_1(y_{j+1}) &= \frac{-f(y_{j+1})c(x_j)}{3.2(\delta y)^2 y_{j+1}} \\
P_1(x_j) &= \frac{-f(y_{j-1})c(x_j)}{3.2(\delta y)^2 y_{j-1}} \\
P_2(x_j) &= \frac{-f(y_{j-2})c(x_j)}{4.3(\delta y)^2 y_{j-2}} \\
P(x_j) &= \frac{f(0)c(x_j)\delta y}{y_{j+2}y_{j+1}y_{j-1}y_{j-2}}
\end{aligned}$$

where we added the minus sign to account for the fact that f is negative for coordinates between y_{-2} and Γ_1. Imposing the symmetry constraints $P_1(y_j) = P_1(x_j)$ we get

$$\frac{f(y_j)c(x_{j-1})}{\cancel{3.2(\delta y)^2} y_j} = \frac{f(y_{j-1})c(x_j)}{\cancel{3.2(\delta y)^2} y_{j-1}}$$

[1]Note that the sequence presented must be reversed to obtain the protocol.

which means

$$c(x_j) = \frac{c_0 f(x_j)}{x_j}$$

where c_0 is a constant. This also entails that $P_2(y_j) = P_2(x_j)$, viz. it satisfies the second symmetry constraint. Finally we can evaluate

$$P(x_j) = \frac{f(0)f(x_j)\delta x}{x_{j+2}x_{j+1}x_j x_{j-1}x_{j-2}} = \frac{c_0 x_0(x_0 - x_j)}{x_j^5}\delta x + \mathcal{O}(\delta x^2)$$

which means that

$$\sum P(x_j) = \frac{1}{2} = \sum \frac{P(x_j)}{x_j} \to \int_{x_0}^{\Gamma} \frac{(x_0 - x)dx}{x^5} = \int_{x_0}^{\Gamma} \frac{(x_0 - x)dx}{x^6}.$$

We can evaluate this as

$$\begin{aligned}
x_0 \int_{x_0}^{\Gamma} \left(\frac{1}{x^5} - \frac{1}{x^6}\right) dx &= \int_{x_0}^{\Gamma} \left(\frac{1}{x^4} - \frac{1}{x^5}\right) dx \\
\left[\frac{1}{4x_0^3} - \frac{1}{5x_0^4}\right] &= \left[\frac{1}{3x_0^3} - \frac{1}{4x_0^4}\right] \\
\left[\frac{1}{4} - \frac{1}{3}\right] &= \left[\frac{1}{5} - \frac{1}{4}\right] \frac{1}{x_0} \\
x_0 \frac{3-4}{\cancel{4}.3} &= \frac{4-5}{5.\cancel{4}} \\
x_0 = \frac{3}{5} \implies \epsilon &= \frac{3}{5} - \frac{1}{2} = \frac{1}{10}.
\end{aligned}$$

4.3.4 Bias $1/10$ Protocol

4.3.4.1 The $3 \to 2$ Move

In this section we introduce as many parameters as possible within the TEF to implement the largest class of $3 \to 2$ moves. However, we use our insight to choose an appropriate basis so that the parameters are small which in turn simplifies the analysis. Henceforth, we use $|g_1\rangle$ instead of $|g_1 g_1\rangle$, for example and similarly $|h_1\rangle$ instead of $|h_1 h_1\rangle$ to avoid cluttering the notation.

Recall that

$$|v\rangle = \frac{\sqrt{p_{g_1}}\,|g_1\rangle + \sqrt{p_{g_2}}\,|g_2\rangle + \sqrt{p_{g_3}}\,|g_3\rangle}{N_g}$$

and let

$$\begin{aligned}
|v_1\rangle &= \frac{\sqrt{p_{g_3}}\,|g_2\rangle - \sqrt{p_{g_2}}\,|g_3\rangle}{N_{v_1}} \\
|v_2\rangle &= \frac{-\frac{(p_{g_2}+p_{g_3})}{\sqrt{p_{g_1}}}\,|g_1\rangle + \sqrt{p_{g_2}}\,|g_2\rangle + \sqrt{p_{g_3}}\,|g_3\rangle}{N_{v_2}}
\end{aligned}$$

where $N_{v_1}^2 = p_{g_3} + p_{g_2}$ and $N_{v_2}^2 = \frac{(p_{g_2}+p_{g_3})^2}{p_{g_1}} + p_{g_2} + p_{g_3}$. Recall that

$$\begin{aligned}
|w\rangle &= \frac{\sqrt{p_{h_1}}\,|h_1\rangle + \sqrt{p_{h_2}}\,|h_2\rangle}{N_h} \\
|w_1\rangle &= \frac{\sqrt{p_{h_2}}\,|h_1\rangle - \sqrt{p_{h_1}}\,|h_2\rangle}{N_h}.
\end{aligned}$$

Now we define

$$\begin{aligned}\left|v_1'\right\rangle &= \cos\theta\left|v_1\right\rangle + \sin\theta\left|v_2\right\rangle \\ \left|v_2'\right\rangle &= \sin\theta\left|v_1\right\rangle - \cos\theta\left|v_2\right\rangle\end{aligned}$$

where we know (from hindsight) that $\cos\theta \approx 1$. The full unitary (which is manifestly unitary) we define to be

$$U = |w\rangle\langle v| + \left(\alpha\left|v_1'\right\rangle + \beta\left|w_1\right\rangle\right)\left\langle v_1'\right| + \left|v_2'\right\rangle\left\langle v_2'\right| + \left(\beta\left|v_1'\right\rangle - \alpha\left|w_1\right\rangle\right)\langle w_1| + |v\rangle\langle w|$$

where $|\alpha|^2+|\beta|^2 = 1$ for $\alpha, \beta \in \mathbb{C}$. There is some freedom in choosing U in the sense that $\alpha|v\rangle+\beta|w_1\rangle$ would also work (then $|v\rangle\langle w| \to |v_1\rangle\langle w|$) because these do not influence the constraint equation. That is what we evaluate now. We need terms of the form $EU|g_i\rangle$ with $E = \mathbb{I}^{\{h_i\}}$. This entails that on the $\{|g_i\rangle\}$ space

$$\begin{aligned}E_hUE_g &= |w\rangle\langle v| + \beta|w_1\rangle\left\langle v_1'\right| \\ &= |w\rangle\langle v| + \beta|w_1\rangle\left(\cos\theta\langle v_1| + \sin\theta\langle v_2|\right).\end{aligned}$$

Consequently we have

$$\begin{aligned}E_hU|g_1\rangle &= \frac{\sqrt{p_{g_1}}}{N_g}|w\rangle + \left[\cos\theta.0 - \sin\theta\frac{p_{g_2}+p_{g_3}}{\sqrt{p_{g_1}}N_{v_2}}\right]\beta|w_1\rangle \\ E_hU|g_2\rangle &= \frac{\sqrt{p_{g_2}}}{N_g}|w\rangle + \left[\cos\theta\frac{\sqrt{p_{g_3}}}{N_{v_1}} + \sin\theta\frac{\sqrt{p_{g_2}}}{N_{v_2}}\right]\beta|w_1\rangle \\ E_hU|g_3\rangle &= \frac{\sqrt{p_{g_3}}}{N_g}|w\rangle + \left[-\cos\theta\frac{\sqrt{p_{g_2}}}{N_{v_1}} + \sin\theta\frac{\sqrt{p_{g_3}}}{N_{v_2}}\right]\beta|w_1\rangle.\end{aligned}$$

Recall that the constraint equation was

$$\sum x_{h_i}|h_i\rangle\langle h_i| - \sum x_{g_i}E_hU|g_i\rangle\langle g_i|U^\dagger E_h \geq 0$$

where the first sum becomes

$$\begin{bmatrix} \langle x_h\rangle & \frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}(x_{h_1}-x_{h_2}) \\ \text{h.c.} & \frac{p_{h_2}x_{h_1}+p_{h_1}x_{h_2}}{N_h^2}\end{bmatrix}$$

in the $|w\rangle, |w_1\rangle$ basis. Since we plan to use the $3 \to 2$ move with one point on the axis, we take $x_{g_1} = 0$. Consequently we need only evaluate

$$\begin{aligned}x_{g_2}E_hU|g_2\rangle\langle g_2|U^\dagger E_h &\doteq x_{g_2}\begin{bmatrix}\frac{p_{g_2}}{N_g^2} & \beta\left(\cos\theta\frac{\sqrt{p_{g_3}p_{g_2}}}{N_gN_{v_1}} + \sin\theta\frac{p_{g_2}}{N_gN_{v_2}}\right) \\ \text{h.c.} & \left(\cos\frac{\sqrt{p_{g_3}}}{N_{v_1}} + \sin\theta\frac{\sqrt{p_{g_2}}}{N_{v_2}}\right)^2|\beta|^2\end{bmatrix} \\ x_{g_3}E_hU|g_3\rangle\langle g_3|U^\dagger E_h &\doteq x_{g_3}\begin{bmatrix}\frac{p_{g_3}}{N_g^2} & \beta\left(-\cos\theta\frac{\sqrt{p_{g_2}p_{g_3}}}{N_gN_{v_1}} + \sin\theta\frac{p_{g_3}}{N_gN_{v_2}}\right) \\ \text{h.c.} & \left(-\cos\frac{\sqrt{p_{g_2}}}{N_{v_1}} + \sin\frac{\sqrt{p_{g_3}}}{N_{v_2}}\right)^2|\beta|^2\end{bmatrix}\end{aligned}$$

which means that the constraint equation becomes

$$\begin{bmatrix} \langle x_h\rangle - \langle x_g\rangle & \frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}(x_{h_1}-x_{h_2}) - \beta\cos\theta\frac{\sqrt{p_{g_2}p_{g_3}}}{N_gN_{v_1}}(x_{g_2}-x_{g_3}) - \beta\sin\theta\langle x_g\rangle\frac{N_g}{N_{v_2}} \\ \text{h.c.} & \frac{p_{h_2}x_{h_1}+p_{h_1}x_{h_2}}{N_h^2} - |\beta|^2\left[\frac{\cos^2\theta}{N_{v_1}^2}(p_{g_3}x_{g_2}+p_{g_2}x_{g_3}) + \frac{\sin^2\theta}{\left(N_{v_2}^2/N_g^2\right)}\langle x_g\rangle + \frac{2\cos\theta\sin\theta\sqrt{p_{g_3}p_{g_2}}}{N_{v_1}N_{v_2}}\left(x_{g_2}-x_{g_3}\right)\right]\end{bmatrix} \geq 0.$$

We already showed that Mochon's transition is average non-decreasing viz. $\langle x_h\rangle - \langle x_g\rangle \geq 0$. We set the off-diagonal elements of the matrix above to zero and show that the second diagonal element, the second eigenvalue therefore, is positive.

Setting the off-diagonal to zero one can obtain θ by solving the quadratic in terms of β although the expression will not be particularly pretty. To establish existence and positivity we need to simplify our expressions.

So far everything was exact even though the basis and techniques were chosen based on experience. Now we claim that $\theta\frac{N_g}{N_{v_2}} = \mathcal{O}(\delta y)$ at most (where $\delta y = \delta x$ is the lattice spacing) and since δy will be taken to be small we can take the small $\theta\frac{N_g}{N_{v_2}}$ limit and to first order in it the constraints become

$$\frac{\frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}(x_{h_1} - x_{h_2}) - \beta\frac{\sqrt{p_{g_2}p_{g_3}}}{N_g N_{v_1}}(x_{g_2} - x_{g_3})}{\beta\langle x_g\rangle} = \theta\frac{N_g}{N_{v_2}} + \mathcal{O}(\delta y^2)$$

and

$$\frac{p_{h_2}x_{h_1} + p_{h_1}x_{h_2}}{N_h^2} - |\beta|^2\left[\frac{p_{g_3}x_{g_2} + p_{g_2}x_{g_3}}{N_{v_1}^2} + 2\theta\frac{N_g}{N_{v_2}}\frac{\sqrt{p_{g_3}p_{g_2}}}{N_g N_{v_1}}(x_{g_2} - x_{g_3})\right] + \mathcal{O}(\delta y^2) \geq 0.$$

If our claim is wrong when we evaluate $\theta\frac{N_g}{N_{v_2}}$ we will get zero order terms but as we show in the following section $\theta\frac{N_g}{N_{v_2}} = 0.\delta y + \mathcal{O}(\delta y^2)$ in fact.

4.3.4.2 Validity of the $3 \to 2$ Move

With respect to Figure 4.3 we have

$$\begin{aligned}
P_2(y_{j+2}) = p_{h_2} &= \frac{-f(y_{j+2})}{4.3\delta y^2 y_{j+2}} \\
P_1(y_{j+1}) = p_{g_3} &= \frac{-f(y_{j+1})}{3.2\delta y^2 y_{j+1}} \\
P_1(x_j) = p_{h_1} &= \frac{-f(y_{j-1})}{3.2\delta y^2 y_{j-1}} \\
P_2(x_j) = p_{g_2} &= \frac{-f(y_{j-2})}{4.3\delta y^2 y_{j-2}} \\
P(x_j) = p_{g_1} &= \frac{f(0)\delta y}{y_{j+2}y_{j+1}y_{j-1}y_{j-2}}
\end{aligned}$$

where we assumed $f(0) > 0$ and $f(y) < 0$ for $y > y_0'$, $y_0' = y_0 + \delta y$. We also scaled by δy to make $P(x_j)$ into a nice density. So far everything is exact. We now convert all expressions to first order in δy. To this end we note

$$\begin{aligned}
f(y_{j+m}) &= f(y_j) + \frac{\partial f}{\partial y}m\delta y + \mathcal{O}(\delta y^2) \\
\frac{1}{y_{j+m}} &= (y_j + m\delta y)^{-1} = \frac{1}{y_j}\left(1 + m\frac{\delta y}{y_j}\right)^{-1} = \frac{1}{y_j} - m\frac{\delta y}{y_j^2} + \mathcal{O}(\delta y^2)
\end{aligned}$$

where $\frac{\partial f}{\partial y}$ refers to $\frac{\partial f(y)}{\partial y}|_{y_j}$. We define and evaluate

$$\begin{aligned}
P_k^m &= \frac{-f(y_{j+m})}{k\delta y^2 y_{j+m}} \\
&= \frac{1}{k\delta y^2}\left[-f(y_j) - \frac{\partial f}{\partial y}m\delta y + \mathcal{O}(\delta y^2)\right]\left[\frac{1}{y_j} - m\frac{\delta y}{y_j^2} + \mathcal{O}(\delta y^2)\right] \\
&= \frac{1}{k\delta y^2}\left[-\frac{f}{y_j} - m\frac{\delta y}{y_j}\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right] \\
&= \frac{1}{ky_j\delta y^2}\left[-f - m\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right]
\end{aligned}$$

where f means $f(y_j)$. In this notation

$$p_{h_2} = P_{12}^2,\ p_{h_1} = P_6^{-1}$$
$$p_{g_2} = P_{12}^{-2},\ p_{g_3} = P_6^1.$$

With an eye at the off-diagonal condition we evaluate

$$P_{k_1}^{m_1} P_{k_2}^{m_2} = \frac{1}{k_1 k_2}\left(\frac{1}{y_j \delta y^2}\right)^2 \left[f^2 + f\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right)(m_1+m_2) + \mathcal{O}(\delta y^2)\right]$$

and

$$P_{k_1}^{m_1} + P_{k_2}^{m_2} = \frac{1}{y_j \delta y^2}\left[-\left(\frac{1}{k_1}+\frac{1}{k_2}\right) f - \left(\frac{m_1}{k_1}+\frac{m_2}{k_2}\right)\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right].$$

We now evaluate

$$\sqrt{p_{h_1} p_{h_2}} = \sqrt{P_{12}^2 P_6^{-1}} = \frac{1}{y_j \delta y^2}\sqrt{\frac{1}{12.6}\left[f^2 + f\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right]}$$

$$N_h^2 = P_{12}^2 + P_6^{-1} = \frac{1}{y_j \delta y^2}\left[-\left(\frac{1}{12}+\frac{1}{6}\right) f - \cancel{\left(\frac{2}{12} - \frac{1}{6}\right)}^{0}\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right]$$
$$= \frac{1}{4 y_j \delta y^2}\left[-f + \mathcal{O}(\delta y^2)\right]$$

and similarly

$$\sqrt{p_{g_2} p_{g_3}} = \sqrt{P_{12}^{-2} P_6^1} = \frac{1}{y_j \delta y^2}\sqrt{\frac{1}{12.6}\left[f^2 - f\delta y\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right) + \mathcal{O}(\delta y^2)\right]}$$

$$N_g^2 = P_{12}^{-2} + P_6^1 + p_{g_1} = \frac{1}{4 y_j \delta y^2}\left[-f + \mathcal{O}(\delta y^2)\right] + \left[\frac{f(0)\delta y}{y_j^4} + \mathcal{O}(\delta y^2)\right]$$
$$= \frac{1}{4 y_j \delta y^2}\left[-f + \mathcal{O}(\delta y^2)\right]$$
$$N_{v_1}^2 = \frac{1}{4 y_j \delta y^2}\left[-f + \mathcal{O}(\delta y^2)\right]$$

where even though it seems like we have neglected p_{g_1} when we take the ratios the meaning of keeping first order in δy would become precise. We can actually take $\beta = 1$ and obtain

$$\theta \frac{N_g}{N_{v_2}} = \frac{4\sqrt{\frac{1}{12.6}}(-3\delta y)\left[f.(\not{1} + \frac{\delta y}{2f}\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right)) - f.(\not{1} - \frac{\delta y}{2f}\left(\frac{\partial f}{\partial y} - \frac{f}{y_j}\right)) + \mathcal{O}(\delta y^2)\right]}{\langle x_g \rangle}$$
$$= 0 + \mathcal{O}(\delta y^2).$$

This shows that to first order the off-diagonal term is zero for $\theta = 0$.

Now we show that the second diagonal element is positive to first order in δy. Using the fact that $\theta \frac{N_g}{N_{v_2}} = \mathcal{O}(\delta y^2)$ we have

$$\frac{p_{h_2} x_{h_1} + p_{h_1} x_{h_2}}{N_h^2} - \frac{p_{g_3} x_{g_2} + p_{g_2} x_{g_3}}{N_{v_1}^2} + \mathcal{O}(\delta y^2) \geq 0$$

as the positivity condition. This becomes

$$
\begin{aligned}
=&\frac{P_{12}^{2}y_{j-1}+P_{6}^{-1}y_{j+2}}{N_h^2}-\frac{P_{6}^{1}y_{j-2}+P_{12}^{-2}y_{j+1}}{N_{v_1}^2}+\mathcal{O}(\delta y^2)\\
=&\left(\frac{4y_j\delta y^2}{-f}\right)\frac{1}{y_j\delta y^2}\\
&\left\{\frac{1}{12}\left[-f-2\delta y\gamma\right](y_j-\delta y)+\frac{1}{6}\left[-f+\gamma\delta y\right](y_j+2\delta y)\right.\\
&\left.-\left(\frac{1}{6}\left[-f-\delta y\gamma\right](y_j-2\delta y)+\frac{1}{12}\left[-f+2\delta y\gamma\right](y_j+\delta y\right)\right\}\\
&+\mathcal{O}(\delta y^2)\\
=&\frac{-2}{3f}\left\{\frac{1}{2}(\cancel{-fy}+f\delta y-2y\delta y\gamma)+(\cancel{-fy}-2f\delta y+y\delta y\gamma)\right.\\
&\left.-\left((\cancel{-fy}+2f\delta y-y\delta y\gamma)+\frac{1}{2}(\cancel{-fy}-f\delta y+2y\delta y\gamma)\right)\right\}\\
&+\mathcal{O}(\delta y^2)\\
=&\frac{-2}{3f}\left\{(f\delta y-2y\delta y\gamma)+2(-2f\delta y+y\delta y\gamma)\right\}+\mathcal{O}(\delta y^2)\\
=&\frac{-2}{3f}\left\{-3f\delta y\right\}+\mathcal{O}(\delta y^2)=2\delta y+\mathcal{O}(\delta y^2)\geq 0
\end{aligned}
$$

where $\gamma=\left(\frac{\partial f}{\partial y}-\frac{f}{y_j}\right)$ and we suppressed the index j in y_j for simplicity. This establishes the validity of the $3\to 2$ transition for a closely spaced lattice.

Note that only the proof of validity was done perturbatively to first order in δy. The unitary itself is known exactly (θ can be obtained by solving the quadratic).
Using $f(y)=(y_0'-y)(\Gamma_1-y)(\Gamma_2-y)$ we can implement the last two moves in Figure 4.3 as they form a $3\to 1$ merge and a $2\to 1$ merge (possibly followed by a raise). The only remaining task is implementing the $2\to 2$ move in the last step because we assumed here that $\sqrt{p_{g_2}}\neq 0$ (else the vectors which we assumed are orthonormal, cease to be so).

4.3.4.3 The $2\to 2$ Move and its validity

We claim that the $2\to 2$ move can be implemented using

$$U=|w\rangle\langle v|+(\alpha|v\rangle+\beta|w_1\rangle)\langle v_1|+|v\rangle\langle w|+(\beta|v\rangle-\alpha|w_1\rangle)\langle w_1|$$

where as before $|\alpha|^2+|\beta|^2=1$,

$$|v\rangle=\frac{1}{N_g}\left(\sqrt{p_{g_1}}|g_1\rangle+\sqrt{p_{g_2}}|g_2\rangle\right),$$

$$|w\rangle=\frac{1}{N_h}\left(\sqrt{p_{h_1}}|h_1\rangle+\sqrt{p_{h_2}}|h_2\rangle\right),$$

$$|v_1\rangle=\frac{1}{N_g}\left(\sqrt{p_{g_2}}|g_1\rangle-\sqrt{p_{g_1}}|g_2\rangle\right),$$

and

$$|w_1\rangle=\frac{1}{N_h}\left(\sqrt{p_{h_2}}|h_1\rangle-\sqrt{p_{h_1}}|h_2\rangle\right).$$

We evaluate the constraint equation using

$$E_h U \left|g_{11}\right\rangle = \frac{\sqrt{p_{g_1}}\left|w\right\rangle + \beta e^{-i\phi_g} e^{i\phi_h}\sqrt{p_{g_2}}\left|w_1\right\rangle}{N_g}$$
$$E_h U \left|g_{22}\right\rangle = \frac{\sqrt{p_{g_2}}\left|w\right\rangle - \beta e^{-i\phi_g} e^{i\phi_h}\sqrt{p_{g_1}}\left|w_1\right\rangle}{N_g}$$

and

$$E_h U \left|g_{11}\right\rangle\left\langle g_{11}\right| U^\dagger E_h = \frac{1}{N_g^2}\begin{array}{c|cc} & \langle w| & \langle w_1| \\ \hline |w\rangle & p_{g_1} & \beta e^{i(\phi_h-\phi_g)}\sqrt{p_{g_2}p_{g_1}} \\ |w_1\rangle & \text{h.c.} & |\beta|^2 p_{g_2} \end{array}$$

(similarly for $L\left|g_{22}\right\rangle\left\langle g_{22}\right| L^\dagger$) as

$$\begin{bmatrix} \langle x_h\rangle - \langle x_g\rangle & \frac{1}{N_g^2}\left[\sqrt{p_{h_1}p_{h_2}}(x_{h_1}-x_{h_2}) - \beta\sqrt{p_{g_1}p_{g_2}}(x_{g_1}-x_{g_2})\right] \\ \text{h.c.} & \frac{1}{N_g^2}\left[p_{h_2}x_{h_1} + p_{h_1}x_{h_2} - |\beta|^2\left(p_{g_2}x_{g_1} + p_{g_1}x_{g_2}\right)\right] \end{bmatrix} \geq 0$$

where we absorbed the phase freedom in β, a free parameter, which will be fixed shortly. We use the same strategy as above and take the first diagonal element to be zero. Our burden would be to first show that

$$\sqrt{\frac{p_{h_1}p_{h_2}}{p_{g_1}p_{g_2}}}\frac{(x_{h_1}-x_{h_2})}{(x_{g_1}-x_{g_2})} = \beta \leq 1$$

and subsequently

$$\frac{1}{N_g^2}\left[p_{h_2}x_{h_1} + p_{h_1}x_{h_2} - |\beta|^2\left(p_{g_2}x_{g_1} + p_{g_1}x_{g_2}\right)\right] \geq 0.$$

What makes this situation special (compared to the $3 \to 2$ merge) is that $f(y_{j-2}) = 0$ which we use to write

$$f(y_{j+k}) = \left.\frac{\partial f}{\partial y}\right|_{y_{j-2}}(k+2)\delta y = -(k+2)\alpha\delta y$$

where

$$\alpha = -\left.\frac{\partial f}{\partial y}\right|_{y_{j-2}} = (\Gamma_1 - y_{j-2})(\Gamma_2 - y_{j-2}).$$

Using the axis situation as depicted in Figure 4.4 we note that

$$p_{h_1} = P_1(x_j) = \frac{-f(y_{j-1})}{3.2\delta y^2 y_{j-1}} = \frac{\alpha + \mathcal{O}(\delta y)}{6\delta y y_j}$$
$$p_{h_2} = P_2(y_{j+2}) = \frac{-f(y_{j+2})}{4.3\delta y^2 y_{j+2}} = \frac{\alpha + \mathcal{O}(\delta y)}{3\delta y y_j}$$
$$x_{h_1} = y_{j-1},\ x_{h_2} = y_{j+2}$$

while

$$p_{g_1} = P(x_j) = \frac{f(0)\delta y}{y_{j+2}y_{j+1}y_{j-1}y_{j-2}} = \frac{f(0)\delta y + \mathcal{O}(\delta y^2)}{y_j^4}$$
$$p_{g_2} = P_1(y_{j+1}) = \frac{-f(y_{j+1})}{3.2\delta y^2 y_{j+1}} = \frac{\alpha + \mathcal{O}(\delta y)}{2\delta y y_j}$$
$$x_{g_1} = 0,\ x_{g_2} = y_{j+1}.$$

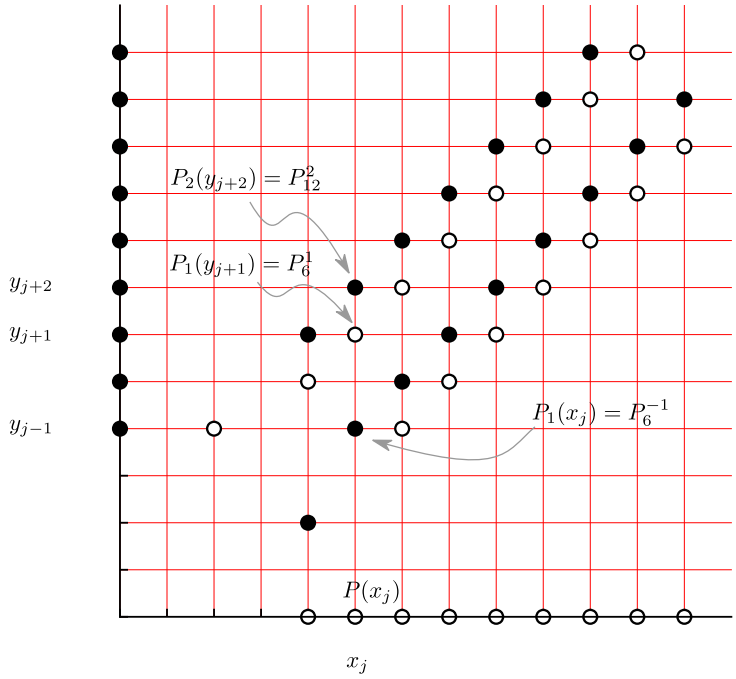


Figure 4.4: First $2 \to 2$ Transition

This entails

$$\beta = \sqrt{\frac{p_{h_1}p_{h_2}}{p_{g_1}p_{g_2}}}\frac{(x_{h_1}-x_{h_2})}{(x_{g_1}-x_{g_2})} = \sqrt{\frac{\alpha^{\cancel{2}} + \mathcal{O}(\delta y)}{\cancel{6}\cancel{3}\delta y^{\cancel{2}}y_j^{\cancel{2}}}\frac{\cancel{2}\delta y y_j^{\cancel{4}}y_j}{\delta y(f(0)\alpha + \mathcal{O}(\delta y))}\frac{(\cancel{3}\delta y)^{\cancel{2}}}{y_j^2 + \cancel{\mathcal{O}(\delta y)}}}$$

$$= \sqrt{\frac{y_0'\alpha + \mathcal{O}(\delta y)}{f(0)}} = \sqrt{\frac{(\Gamma_1 - y_{j-2})(\Gamma_2 - y_{j-2}) + \mathcal{O}(\delta y)}{\Gamma_1\Gamma_2}} \leq 1$$

where we used $f(0) = y_0'\Gamma_1\Gamma_2$ and assumed δy is small compared Γs (which is the case) for the inequality in the last step to hold.

The second condition can be proven similarly

$$\frac{1}{N_g^2}\left[p_{h_2}x_{h_1} + p_{h_1}x_{h_2} - |\beta|^2\left(p_{g_2}x_{g_1} + p_{g_1}x_{g_2}\right)\right]$$

$$\geq \frac{1}{N_g^2}\left[p_{h_2}x_{h_1} + p_{h_1}x_{h_2} - p_{g_2}x_{g_1}\right]$$

$$= \frac{1}{N_g^2}\left[\frac{\alpha + \mathcal{O}(\delta y)}{3\delta y y_j}y_{j-1} + \frac{\alpha + \mathcal{O}(\delta y)}{6\delta y y_j}y_{j+2} - \frac{f(0)\delta y + \mathcal{O}(\delta y^2)}{y_j^4}y_{j+1}\right]$$

$$= \frac{1}{3\delta y N_g^2}\left[(\alpha + \mathcal{O}(\delta y))\left(\frac{3}{2}\right) - \underbrace{\frac{f(0)\delta y^2 + \mathcal{O}(\delta y^3)}{y_j^3}}_{\in\mathcal{O}(\delta y^2)}\right]$$

$$= \frac{1}{2\delta y N_g^2}\left[(\Gamma_1 - y_{j-2})(\Gamma_2 - y_{j-2}) + \mathcal{O}(\delta y)\right] \geq 0$$

where the last step holds for δy small enough.

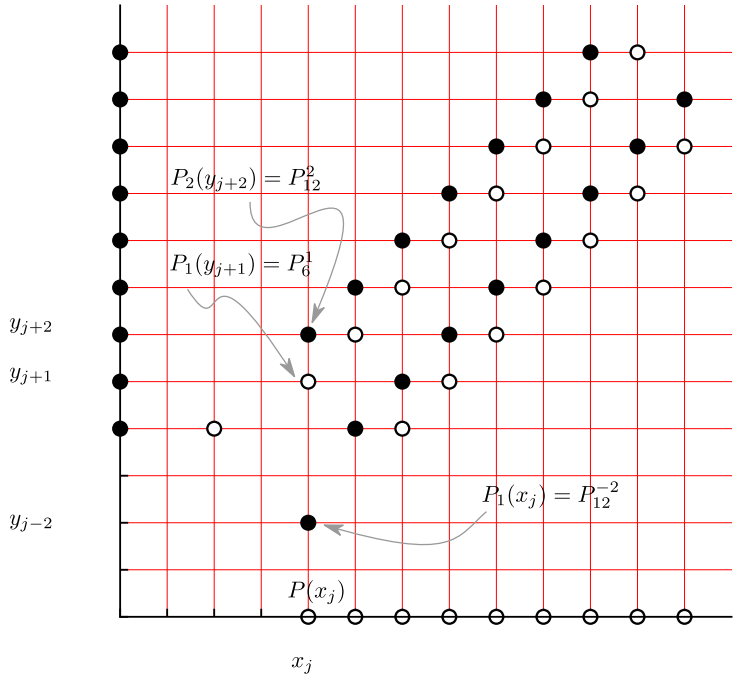


Figure 4.5: Final $2 \to 2$ Transition.

The $2 \to 2$ move corresponding to the leftmost (see Figure 4.5) and bottommost set of points can be shown to be implementable similarly.

5

Chapter

Approaching bias $1/(4k+2)$

While we succeeded at constructing the unitaries involved in the bias $1/10$ protocol, we did not follow a systematic method. In this chapter, we construct the unitaries corresponding to Mochon's f-assignments (see Lemma 45). These, together with TEF and the previous results, allow us to construct explicit WCF protocols with vanishing bias, to wit, bias approaching $1/(4k+2)$ for arbitrary integers $k > 1$.

Notation. We follow these notations.

- We use, for a Hermitian matrix $A = \sum_i a_i |i\rangle\langle i|$ (its spectral decomposition, including zero eigenvalues), we define a pseudo-inverse or a generalised of A as $A^{\dashv} := \sum_{i:|a_i|>0} a_i^{-1} |i\rangle\langle i|$, so that, e.g. if $A = \begin{bmatrix} 0 & & \\ & 2 & \\ & & 3 \end{bmatrix}$ then $A^{\dashv} = \begin{bmatrix} 0 & & \\ & 1/2 & \\ & & 1/3 \end{bmatrix}$.
- We follow the convention (unless otherwise stated) of writing functions t with finite support in the following two ways: (1) as $t = \sum_{i=}^{n} p_i [\![x_i]\!]$ where we assume $p_i > 0$ for all $i \in \{1, 2 \dots n\}$ and that $x_i \neq x_j$ for $i \neq j$ and (2) as $t = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ where p_{h_i} and p_{g_i} are strictly positive and x_{h_i} and x_{g_i} are all distinct.
- Motivated by Theorem 10 and Theorem 75, given a function (with finite support) $t = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ and an orthonormal basis $\{|g_1\rangle, |g_2\rangle \dots |g_{n_g}\rangle, |h_1\rangle, |h_2\rangle \dots |h_{n_h}\rangle\}$ we say that an orthogonal matrix O solves t if O satisfies the following: $O|v\rangle = |w\rangle$ and $X_h \geq E_h O X_g O^T E_h$ where $|v\rangle = \sum_{i=1}^{n_g} \sqrt{p_{g_i}} |g_i\rangle$, $|w\rangle = \sum_{i=1}^{n_h} \sqrt{p_{h_i}} |h_i\rangle$, $X_h = \sum_{i=1}^{n_h} x_{h_i} |h_i\rangle\langle h_i|$, $X_g = \sum_{i=1}^{n_g} x_{g_i} |g_i\rangle\langle g_i|$ and $E_h = \sum_{i=1}^{n_h} |h_i\rangle\langle h_i|$.

§ 5.1 Mochon's Assignments

We have seen Mochon's assignment in Chapter 2 (see Lemma 45) and Chapter 4 (see Definition 4.3.1.1) already. We tailor it to our purpose and restate it.

Definition 80 (Mochon's f-assignment, f_0-assignment, balanced assignment)**.** Given a set of real numbers $0 \leq x_1 < x_2 \cdots < x_n$ and a polynomial of degree at most $n-2$ satisfying $f(-\lambda) \geq 0$ for all $\lambda \geq 0$, *Mochon's f-assignment* is given by the function

$$t = \sum_{i=1}^{n} \underbrace{\frac{-f(x_i)}{\prod_{j\neq i}(x_j - x_i)}}_{:=p_i} [\![x_i]\!] = h - g,$$

(up to a positive multiplicative factor) where h contains the positive part of t and g the negative part (without any common support), viz. $h = \sum_{i:p_i>0} p_i [\![x_i]\!]$ and $g = \sum_{i:p_i<0} (-p_i) [\![x_i]\!]$.

- When f is a monomial, viz. has the form $f(x) = x^k$, we call the assignment a *monomial-assignment* or an m_k*-assignment*. We also refer to an m_0-assignment as an f_0-assignment.
- We say an assignment is *balanced* if the number of points with negative weights, $p_i < 0$, equals the number of points with positive weights, $p_i > 0$. We say an assignment is *unbalanced* if it is not balanced.

It is easy to see that Mochon's f_0-assignment starts with a point that has a negative weight, regardless of the number of points used to define the assignment. Thereafter, the sign alternates. With this as the base structure, working out the signs of the weights for m-assignments is facilitated. These considerations become relevant when we construct analytic solutions. However, the only mathematical property of Mochon's assignments which is needed to find an analytic solution, turns out to be the following (which jointly restates Lemma 77 and Lemma 78).

Lemma 81. *Let* $t = \sum_{i=1}^{n} p_i [\![x_i]\!]$ *be Mochon's* f_0*-assignment for a set of real numbers* $0 \leq x_1 < x_2 \cdots < x_n$*. Then for* $0 \leq k \leq n-2$*,*

$$\left\langle x^k \right\rangle = 0,$$

and

$$\left\langle x^{n-1} \right\rangle > 0,$$

where $\left\langle x^l \right\rangle = \sum_{i=1}^{n} p_i \, (x_i)^l$.

§ 5.2 Equivalence to Monomial Assignments

While Mochon's f-assignment is a valid function for all polynomials f satisfying the conditions in Definition 80, in the remainder of this chapter, we restrict to polynomials f with real roots (which to be consistent with Definition 80 must in fact additionally be non-negative real roots).

In this section we observe that Mochon's f-assignments can be, almost trivially, expressed as a sum of monomial assignments. In the following sections, we give the unitaries corresponding to these monomial assignments.

We briefly outline why it suffices to implement the valid functions whose sum is the f-function—the valid function corresponding to an f-assignment—to implement the f-function itself. The difficulty is that the valid functions which constitute the sum might have assigned a negative weight to a point to which a positive weight is assigned by the f-assignment. However, we have faced a similar difficulty while transforming a TIPG into a TDPG. It was handled using the ''catalyst state'' (following Kitaev/Mochon and Aharanov et al) in the proof of Theorem 39 in Chapter 2. The basic idea there was to introduce a small negative weight and apply the valid functions by appropriately scaling them down (so that there is sufficient weight at the locations of negative weight) repeatedly to have the same effect as having applied the unscaled valid function. The small negative weight could be made arbitrarily small at the expense of communication rounds, thereby having a vanishing effect on the bias. This technique also lets us apply the valid functions which constitute the sum instead of applying the given f function. This should be straight forward; we do not make it more precise in this thesis.

We first give the obvious decomposition of an f-assignment into monomials which also happens to be quite general. Another decomposition is given in Section C.1 of the Appendix.

Lemma 82 (f-assignment to sum of monomials)**.** *Consider a set of real coordinates satisfying* $0 \leq x_1 < x_2 \cdots < x_n$ *and let* $f(x) = (r_1 - x)(r_2 - x)\dots(r_k - x)$ *where* $k \leq n-2$*. Let* $t = \sum_{i=1}^{n} p_i \, [\![x_i]\!]$ *be the corresponding Mochon's* f*-assignment (justifying the restriction on* k*). Then*

$$t = \sum_{l=0}^{k} \alpha_l \left(\sum_{i=1}^{n} \frac{-(-x_i)^l}{\prod_{j \neq i}(x_j - x_i)} \, [\![x_i]\!] \right)$$

where $\alpha_l \geq 0$ *(more precisely, it is the coefficient of* $(-x)^l$ *in* $f(x)$*).*

5.2.1 The origin can be shifted

Later, we will have to use matrix inverses. Having a coordinate equal to zero, therefore, breaks our argument. However, the following lemma tells us that the solution to Mochon's f-assignment is invariant under a shift of the origin.

Lemma 83. *Consider a set of real numbers/coordinates satisfying* $0 \leq x_1 < x_2 \cdots < x_n$ *and let* $f(x) = (a_1 - x)(a_2 - x)\dots(a_k - x)$ *where* $k \leq n-2$ *and the roots* $\{a_i\}_{i=1}^{k}$ *of* f *are non-negative. Let* $t = \sum_{i=1}^{n} p_i \, [\![x_i]\!]$ *be the corresponding Mochon's* f*-assignment. Consider a set of real coordinates satisfying* $0 < x_1 + c < x_2 + c \cdots < x_n + c$ *where* $c > 0$ *and let* $f'(x) = (a_1 + c - x)(a_2 + c - x)\dots(a_k + c - x)$*. Let* $t' = \sum_{i=1}^{n} p'_i \, [\![x'_i]\!]$ *be the corresponding Mochon's* f*-assignment with* $x'_i := x_i + c$*. The solution to* t *and to* t' *are the same.*

Proof sketch. Note that $p'_i = p_i$ as the c's cancel. As stated at the beginning of the chapter, we write $t = \sum_{i=1}^{n_h} p_{h_i} \, [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} p_{g_i} \, [\![x_{g_i}]\!]$ and define $X_h := \sum_{i=1}^{n_h} x_{h_i} \, |h_i\rangle$, $X_g := \sum_{i=1}^{n_g} x_{g_i} \, |g_i\rangle$. If t is solved by O then we must have $X_h \geq E_h O X_g O^T E_h$. We show that $X_h + c\mathbb{I}_h \geq E_h O (X_g + c\mathbb{I}_g) O^T E_h$ where $\mathbb{I}_h := \sum_{i=1}^{n_h} |h_i\rangle \langle h_i|$ and $\mathbb{I}_g := \sum_{i=1}^{n_g} |g_i\rangle \langle g_i|$. Together with the observation that $p'_i = p_i$, this establishes that O also solves t'. Since c is an arbitrary real number, it follows that O solves t if and only if it solves t'.

We now establish $X_h \geq E_h O X_g O^T E_h \iff X_h + c\mathbb{I}_h \geq E_h O (X_g + c\mathbb{I}_g) O^T E_h$. Observe

$$\begin{aligned}
& X_h \geq E_h O X_g O^T E_h \\
\iff \; & E_h (X_h - O X_g O^T) E_h \geq 0 && \because X_h = E_h X_h E_h \\
\iff \; & E_h (X_h + c\mathbb{I}_{hg} - O(X_g - c\mathbb{I}_{hg}) O^T) E_h \geq 0 \\
\iff \; & X_h + c\mathbb{I}_h \geq E_h O (X_g + c\mathbb{I}_{hg}) O^T E_h, && \text{where } \mathbb{I}_{hg} := \mathbb{I}.
\end{aligned}$$

Further,

$$\begin{aligned}
& X_g + c\mathbb{I}_{hg} \geq X_g + c\mathbb{I}_g \\
\implies \; & E_h O (X_g + c\mathbb{I}_{hg}) O^T E_h \geq E_h O (X_g + c\mathbb{I}_g) O^T E_h
\end{aligned}$$

which together yield

$$X_h \geq E_h O X_g O^T E_h \iff X_h + c\mathbb{I}_h \geq E_h O (X_g + c\mathbb{I}_g) O^T E_h.$$

□

Now it only remains to solve monomial assignments, a task which occupies us for the remaining chapter.

§ 5.3 f_0 Unitary | Solution to Mochon's f-assignment

We begin with the unitaries corresponding to the simplest f-assignment, the f_0 assignment (i.e. $f(x) = (-x)^0$). We first look at the balanced case, where the number of points involved, $2n$, is even. This corresponds to an $n \to n$ transition.

5.3.1 The Balanced Case

Proposition 84 (Solution to balanced f_0 assignments)**.** *Let*

- *an f_0-assignment over $\{x_1, x_2 \dots x_{2n}\}$ be given by*

$$t = \sum_{i=1}^{n} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} p_{g_i} [\![x_{g_i}]\!]$$

- *$\{|h_1\rangle, |h_2\rangle \dots |h_n\rangle, |g_1\rangle, |g_2\rangle \dots |g_n\rangle\}$ be an orthonormal basis (see Theorem 75 with $|h_i h_i\rangle \leftrightarrow |h_i\rangle$ and $|g_i g_i\rangle \leftrightarrow |g_i\rangle$),*
- *finally*

$$X_h := \sum_{i=1}^{n} x_{h_i} |h_i\rangle \langle h_i| \doteq diag(x_{h_1}, x_{h_2} \dots x_{h_n}, \underbrace{0, 0 \dots 0}_{n\text{-zeros}}),$$

$$X_g := \sum_{i=1}^{n} x_{g_i} |g_i\rangle \langle g_i| \doteq diag(\underbrace{0, 0, \dots 0}_{n\text{-zeros}}, x_{g_1}, x_{g_2} \dots x_{g_n}),$$

$$|w\rangle := \sum_{i=1}^{n} \sqrt{p_{h_i}} |h_i\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots \sqrt{p_{h_n}}, \underbrace{0, 0 \dots 0}_{n\text{-zeros}})^T$$

and

$$|v\rangle := \sum_{i=1}^{n} \sqrt{p_{g_i}} |g_i\rangle \doteq (\underbrace{0, 0 \dots 0}_{n\text{-zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots \sqrt{p_{g_n}})^T.$$

Then,

$$O := \sum_{i=0}^{n-1} \left(\frac{\Pi^{\perp}_{h_{i-1}} (X_h)^i |w\rangle \langle v| (X_g)^i \Pi^{\perp}_{g_{i-1}}}{\sqrt{c_{h_i} c_{g_i}}} + h.c. \right)$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad \text{and} \quad O|v\rangle = |w\rangle$$

where $E_h := \sum_{i=1}^{n} |h_i\rangle \langle h_i|$, $\Pi^{\perp}_{h_{-1}} = \Pi^{\perp}_{g_{-1}} = \mathbb{I}$,

$$\Pi^{\perp}_{h_i} := \text{projector orthogonal to span}\{(X_h)^i |w\rangle, (X_h)^{i-1} |w\rangle, \dots |w\rangle\},$$

$c_{h_i} := \langle w| (X_h)^i \Pi^{\perp}_{h_{i-1}} (X_h)^i |w\rangle$ *and analogously*

$$\Pi^{\perp}_{g_i} := \text{projector orthogonal to span}\{(X_g)^i |v\rangle, (X_g)^{i-1} |v\rangle, \dots |v\rangle\},$$

$c_{g_i} := \langle v| (X_g)^i \Pi^{\perp}_{g_{i-1}} (X_g)^i |v\rangle$.

Proof. We use

$$\left\langle x^k \right\rangle = 0 \tag{5.1}$$

for $k \in \{0, 1, 2 \ldots, 2n-2\}$ and that

$$\left\langle x^{2n-1} \right\rangle > 0. \tag{5.2}$$

We define the basis of interest here, essentially using the Gram-Schmidt procedure. Let

$$\begin{aligned}
|w_0\rangle &:= |w\rangle \\
|w_1\rangle &:= \frac{(\mathbb{I} - |w_0\rangle\langle w_0|)(X_h)|w\rangle}{\sqrt{c_{h_1}}} \\
|w_2\rangle &:= \frac{(\mathbb{I} - |w_1\rangle\langle w_1| - |w_0\rangle\langle w_0|)(X_h)|w\rangle}{\sqrt{c_{h_2}}} \\
&\vdots \\
|w_k\rangle &:= \frac{\left(\mathbb{I} - \sum_{i=0}^{k-1} |w_i\rangle\langle w_i|\right)(X_h)^k |w\rangle}{\sqrt{c_{h_k}}}.
\end{aligned} \tag{5.3}$$

We indicate the term with the highest power of X_h appearing in $|w_k\rangle$ by

$$\mathcal{M}(|w_k\rangle) = \left\langle x_h^{2k} \right\rangle \cdot (X_h)^k |w\rangle$$

where the scalar in the numerator represents the dependence on the highest power of x_h (appearing as $\langle x_h^l \rangle$) in $|w_k\rangle$. For instance, here the $\langle x_h^{2k} \rangle$ factor comes from $\sqrt{c_{h_k}}$. Note that the projectors can be expressed in terms of these vectors more concisely,

$$\Pi_{h_i} := \mathbb{I} - \Pi_{h_i}^{\perp} = \sum_{j=0}^{i} |w_j\rangle\langle w_j|.$$

It also follows that O can be re-written as

$$O = \sum_{j=0}^{n-1} (|w_j\rangle\langle v_j| + |v_j\rangle\langle w_j|)$$

where $|v_j\rangle$ is analogously defined (by replacing hs with gs). It is evident that $O|v\rangle = |w\rangle$. We set $D = X_h - E_h O X_g O^T E_h$ and note that $\langle v_j| D |v_i\rangle = 0$ (because $X_h|v_i\rangle = 0$ and $E_h|v_i\rangle = 0$[1]). We assert that it has the following rank-1 form

$$D = \begin{bmatrix} 0 & 0 & \ldots & & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & \ddots & & \\ & & & 0 & 0 \\ 0 & \ldots & & 0 & \langle w_{n-1}| D |w_{n-1}\rangle \end{bmatrix}$$

in the $(|w_0\rangle, |w_1\rangle, \ldots |w_{n-1}\rangle)$ basis, together with $\langle w_{n-1}| D |w_{n-1}\rangle > 0$. To see this, we simply compute

$$\begin{aligned}
\langle w_i| D |w_j\rangle &= \langle w_i| X_h |w_j\rangle - \langle w_i| O X_g O^T |w_j\rangle \\
&= \langle w_i| X_h |w_j\rangle - \langle v_i| X_g |v_j\rangle.
\end{aligned}$$

[1]The conclusion holds even without the projector as O maps $\text{span}(|v_1\rangle, |v_2\rangle, \ldots |v_n\rangle)$ to $\text{span}(|w_1\rangle, |w_2\rangle \ldots |w_n\rangle)$ on which X_g has no support.

For (i,j) for any $0 \leq i,j \leq n-1$ except for the case where both $i=j=n-1$, the two terms are the same. This is because the term with the highest possible power l (of $\langle x^l \rangle$) in $\langle w_i| X_h |w_j\rangle$ can be deduced by observing

$$\mathcal{M}(\langle w_i|)X_h\mathcal{M}(|w_j\rangle) = \left\langle x_h^{2i} \right\rangle \cdot \left\langle x_h^{2j} \right\rangle \cdot \left\langle x_h^{i+j+1} \right\rangle. \tag{5.4}$$

For the analogous expression with gs to be the same, we must have $2i, 2j$ and $i+j+1$ less than or equal to $2n-2$. The first two are always satisfied (for $0 \leq i,j \leq n-1$). The last can only be violated when $i=j=n-1$. This establishes that the matrix has the asserted form.

To prove the positivity of $\langle w_{n-1}| D |w_{n-1}\rangle$, consider $\langle w_{n-1}| X_h |w_{n-1}\rangle$ and $\langle v_{n-1}| X_g |v_{n-1}\rangle$. When these terms are expanded in powers of $\langle x_h^k \rangle$ and $\langle x_g^k \rangle$ respectively, only terms with $k > 2n-2$ would remain; the others would get cancelled due to Equation (5.1). It follows that (using Equation (5.3))

$$\langle w_{n-1}| D |w_{n-1}\rangle = \frac{1}{c_{h_{n-1}}} \langle w| (X_h)^{2n-2+1} |w\rangle - \frac{1}{c_{g_{n-1}}} \langle v| (X_g)^{2n-2+1} |v\rangle$$

and it is not hard to see that $c_{h_{n-1}} = c_{h_{n-1}}(\langle x_h^{2n-2} \rangle, \langle x_h^{2n-3} \rangle, \ldots, \langle x_h^1 \rangle)$ does not depend on $\langle x_h^{2n-1} \rangle$ (we proceed analogously for $c_{g_{n-1}}$). Further, $c_{h_{n-1}} = c_{g_{n-1}} =: c_{n-1}$. We thus have

$$\langle w_{n-1}| D |w_{n-1}\rangle = \frac{\langle x_h^{2n-1} \rangle}{c_{n-1}} > 0$$

using Equation (5.2). Thus, $X_h - E_h O X_g O^T E_h \geq 0$.

Note that in the analysis, we assumed that $\text{span}\{|w\rangle, X_h |w\rangle, X_h^2 |w\rangle, \ldots, X_h^n |w\rangle\}$ equals $\text{span}\{|h_1\rangle, |h_2\rangle \ldots |h_n\rangle\}$ which is justified by Lemma 85. □

Lemma 85. *Consider an n-dimensional vector space. Given a diagonal matrix $X = diag(x_1, x_2 \ldots x_n)$ and a vector $|c\rangle = (c_1, c_2 \ldots, c_n)$ where all the x_is are distinct and all the c_i are non-zero, the vectors $|c\rangle, X|c\rangle, \ldots X^{n-1}|c\rangle$ span the vector space.*

Proof. We write the vectors as

$$|\tilde{w}_i\rangle = X^{i-1}|c\rangle = \begin{bmatrix} x_1^{i-1}c_1 \\ x_2^{i-1}c_2 \\ \vdots \\ x_n^{i-1}c_n \end{bmatrix}.$$

We show that the set of vectors are linearly independent, which is equivalent to showing that the determinant of the matrix containing the vectors as rows (or equivalently as columns) is non-zero, i.e.

$$\det\left(\underbrace{\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & & x_n \\ x_1^2 & x_2^2 & & x_n^2 \\ \vdots & & \ddots & \\ x_1^{n-1} & x_2^{n-1} & \ldots & x_n^{n-1} \end{bmatrix}}_{:=\tilde{X}} \begin{bmatrix} c_1 & & & \\ & c_2 & & \\ & & \ddots & \\ & & & c_n \end{bmatrix} \right) = c_1 \cdot c_2 \cdot \ldots c_n \cdot \det \tilde{X}$$

is non-zero. To see this, we note that $\tilde{X}$ is the so-called Vandermonde matrix (restricted to being a square matrix) and its determinant, known as the Vandermonde determinant, is $\det(\tilde{X}) = \prod_{1 \leq i \leq j \leq n}(x_j - x_i) \neq 0$ as x_is are distinct. As c_is are all non-negative, this concludes the proof. □

5.3.2 The Unbalanced Case

We can now consider unbalanced f_0 assignments. Before proceeding, let us try to understand the result we just proved slightly better and see where it fails. We could write $D_{ij} = \langle w_i| D |w_j\rangle$ and note that the maximum power l which appears as $\left\langle x^l_{g/h} \right\rangle$ is given by $\max\{2i, 2j, i+j+1\}$. This yields a matrix with each term depending on the power as

$$D = \begin{bmatrix} D_{00}(\langle x\rangle) & & & & & \\ D_{10}(\langle x^2\rangle, \dots) & D_{11}(\langle x^3\rangle, \dots) & & & & \text{h.c.} \\ D_{20}(\langle x^4\rangle, \dots) & D_{21}(\langle x^4\rangle, \dots) & D_{22}(\langle x^5\rangle, \dots) & & & \\ D_{30}(\langle x^6\rangle, \dots) & D_{31}(\langle x^6\rangle, \dots) & D_{32}(\langle x^6\rangle, \dots) & D_{33}(\langle x^7\rangle, \dots) & & \\ D_{40}(\langle x^8\rangle, \dots) & D_{41}(\langle x^8\rangle, \dots) & D_{42}(\langle x^8\rangle, \dots) & D_{43}(\langle x^8\rangle, \dots) & D_{44}(\langle x^9\rangle, \dots) & \\ & & & & & \ddots \end{bmatrix}.$$

For brevity, we represent this dependence as

$$\mathcal{M}(D) = \begin{bmatrix} \langle x\rangle & & & \\ \langle x^2\rangle & \langle x^3\rangle & & \\ \langle x^4\rangle & \langle x^4\rangle & \langle x^5\rangle & \\ & & & \ddots \end{bmatrix}.$$

Consider the balanced f_0 case over $\{x_1, x_2, x_3, x_4\}$. In this case $\langle x\rangle = \langle x^2\rangle = 0$ and $\langle x^3\rangle > 0$. This is a two dimensional case. It thus follows that

$$\mathcal{M}(D) = \begin{bmatrix} 0 & 0 \\ 0 & \langle x^3\rangle \end{bmatrix} \geq 0.$$

If we now try to use the same procedure for an f_0 assignment over $\{x_1, x_2 \dots x_5\}$, we'll have $\langle x\rangle = \langle x^2\rangle = \langle x^3\rangle = 0$ and $\langle x^4\rangle > 0$. If we try to solve in three dimensions, we would obtain

$$\mathcal{M}(D) = \begin{bmatrix} 0 & 0 & \langle x^4\rangle \\ 0 & 0 & \langle x^4\rangle \\ \langle x^4\rangle & \langle x^4\rangle & \langle x^5\rangle \end{bmatrix} \tag{5.5}$$

which does not seem to work directly. It turns out that the projector that was present in the TEF constraint, gets rid of the troublesome part and yields a zero matrix. We see it in this example first and then generalise it. The unbalanced assignment takes three points to two points. We define $X_h := \text{diag}(x_{h_1}, x_{h_2}, 0, 0, 0)$, $|w\rangle = (\sqrt{p_{h_1}}, \sqrt{p_{h_2}}, 0, 0, 0)$ along with

$$\begin{aligned} |w_0\rangle &:= |w\rangle \\ |w_1\rangle &:= (\mathbb{I} - |w_0\rangle \langle w_0|) X_h |w_0\rangle . \end{aligned}$$

We can write $E_h = \sum_{i=0}^{1} |w_i\rangle \langle w_i|$ and have the same unitary as before, except that we leave $|v_2\rangle$ unchanged, i.e. $O = \sum_{i=0}^{1} |w_i\rangle \langle v_i| + |v_2\rangle \langle v_2|$. We can now show that $D' = X_h - E_h O X_g O^T E_h \geq 0$ because every vector in $|\psi\rangle \in \text{span}\{|v_0\rangle, |v_1\rangle, |v_2\rangle\}$ satisfies $D'|\psi\rangle = 0$ (as $X_h |\psi\rangle = 0$ and $E_h |\psi\rangle = 0$). This entails it suffices to restrict to a 2×2 matrix in $\text{span}\{|w_0\rangle, |w_1\rangle\}$. But this we already know is zero (from Equation (5.5)), hence $D' = 0$. This might seem surprising at first but recall that even for the merge the projected matrix condition becomes a scalar condition (so a 1×1 matrix) which is saturated, i.e. set to zero.

Proposition 86 (Solution to unbalanced f_0-assignments)**.** *Let*

- *an f_0-assignment over $0 < x_1 < x_2 \cdots < x_{2n-1}$ be given by*

$$t = \sum_{i=1}^{n-1} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} p_{g_i} [\![x_{g_i}]\!],$$

- $\{|h_1\rangle, |h_2\rangle \ldots |h_{n-1}\rangle, |g_1\rangle, |g_2\rangle \ldots |g_n\rangle\}$ *be an orthonormal basis*
- *finally*

$$X_h := \sum_{i=1}^{n-1} x_{h_i} |h_i\rangle \langle h_i| \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_{n-1}}, \underbrace{0, 0, \ldots 0}_{n \text{ zeros}}),$$

$$X_g := \sum_{i=1}^{n} x_{g_i} |g_i\rangle \langle g_i| \doteq diag(\underbrace{0, 0, \ldots 0}_{n-1 \text{ zeros}}, x_{g_1}, x_{g_2} \ldots x_{g_{n-1}}, x_{g_n}),$$

$$|w\rangle := \sum_{i=1}^{n-1} \sqrt{p_{h_i}} |h_i\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}}, \ldots \sqrt{p_{h_{n-1}}}, \underbrace{0, 0 \ldots 0}_{n \text{ zeros}})^T,$$

and

$$|v\rangle := \sum_{i=1}^{n} \sqrt{p_{g_i}} |g_i\rangle \doteq (\underbrace{0, 0, \ldots 0}_{n-1 \text{ zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_{n-1}}}, \sqrt{p_{g_n}})^T$$

- *and $E_h := \sum_{i=1}^{n-1} |h_i\rangle \langle h_i|$.*

Then,

$$O := \left(\sum_{i=0}^{n-2} \frac{\Pi^{\perp}_{h_{i-1}} (X_h)^i |w\rangle \langle v| (X_g)^i \Pi^{\perp}_{g_{i-1}}}{\sqrt{c_{h_i} c_{g_i}}} + h.c. \right) + \frac{\Pi^{\perp}_{g_{n-2}} (X_g)^{n-1} |v\rangle \langle v| (X_g)^{n-1} \Pi^{\perp}_{g_{n-2}}}{c_{g_i}}$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad \text{and} \quad E_h O |v\rangle = |w\rangle$$

where $\Pi^{\perp}_{h_{-1}} = \Pi^{\perp}_{g_{-1}} = \mathbb{I}$,

$$\Pi^{\perp}_{h_i} := \text{projector orthogonal to span}\{(X_h)^i |w\rangle, (X_h)^{i-1} |w\rangle, \ldots |w\rangle\},$$

$c_{h_i} := \langle w| (X_h)^i \Pi^{\perp}_{h_{i-1}} (X_h)^i |w\rangle$ *and analogously*

$$\Pi^{\perp}_{g_i} := \text{projector orthogonal to span}\{(X_g)^i |v\rangle, (X_g)^{i-1} |v\rangle, \ldots |v\rangle\},$$

$c_{g_i} := \langle v| (X_g)^i \Pi^{\perp}_{g_{i-1}} (X_g)^i |v\rangle$.

Proof. We will again, use

$$\left\langle x^k \right\rangle = 0 \tag{5.6}$$

but this time, $k \in \{0, 1, \ldots 2n-3\}$ and

$$\left\langle x^{2n-2} \right\rangle > 0.$$

We define the basis, almost exactly as before, we set $|w_0\rangle := |w\rangle$ and for each integer k satisfying $0 \leq k \leq n-2$ we have

$$|w_k\rangle := \frac{\Pi^{\perp}_{h_{k-1}} (X_h)^k |w\rangle}{\sqrt{c_{h_k}}} = \frac{\left(\mathbb{I} - \sum_{i=0}^{k-1} |w_i\rangle \langle w_i|\right) (X_h)^k |w\rangle}{\sqrt{c_{h_k}}}.$$

We define $|v_0\rangle := |v\rangle$ and for each integer satisfying $0 \leq k \leq n-1$ we have

$$|v_k\rangle := \frac{\Pi^{\perp}_{g_{k-1}} (X_g)^k |v\rangle}{\sqrt{c_{g_k}}} = \frac{\left(\mathbb{I} - \sum_{i=0}^{k-1} |v_i\rangle \langle v_i|\right) (X_g)^k |v\rangle}{\sqrt{c_{h_k}}}.$$

Note that this means $O = \sum_{i=0}^{n-2} (|w_i\rangle \langle v_i| + |v_i\rangle \langle w_i|) + |v_n\rangle \langle v_n|$ and so $E_h O |v\rangle = |w\rangle$ follows directly. Also, to establish $D := X_h - E_h O X_g O^T E_h \geq 0$, note that it suffices to show that $\langle w_i| D |w_j\rangle \geq 0$ for integers i, j satisfying $0 \leq i, j \leq n-2$. This is because, as we saw in the previous case, $D |v_i\rangle = 0$ as $X_h |v_i\rangle = 0$ and $E_h |v_i\rangle = 0$. As before, we indicate the term with the highest power of X_h appearing in $|w_k\rangle$, for k in $\{0, 1 \ldots n-2\}$, by

$$\mathcal{M}(|w_k\rangle) = \left\langle x_h^{2k} \right\rangle \cdot (X_h)^k |w\rangle$$

and analogously, the highest power of X_g appearing in $|v_k\rangle$ for k in $\{0, 1, \ldots n-2\}$, by

$$\mathcal{M}(|v_k\rangle) = \left\langle x_g^{2k} \right\rangle \cdot (X_g)^k |v\rangle .$$

Again, the highest power l of $\langle x^l \rangle$ that appears in $\langle w_i| D |w_j\rangle$ is $\max\{2j, 2i, i+j+1\}$ which can be deduced by evaluating

$$\mathcal{M}(\langle w_i|) X_h \mathcal{M}(|w_j\rangle) = \left\langle x_h^{2j} \right\rangle \cdot \left\langle x_h^{2i} \right\rangle \cdot \left\langle x_h^{i+j+1} \right\rangle$$

and similarly

$$\mathcal{M}(\langle v_i|) E_h O X_g O E_h \mathcal{M}(|v_i\rangle) = \left\langle x_g^{2j} \right\rangle \cdot \left\langle x_g^{2i} \right\rangle \cdot \left\langle x_g^{i+j+1} \right\rangle .$$

The highest possible power is obtained when $i = j = n-2$. This yields $2n-3$ and thus, using Equation (5.6), we conclude that $\langle w_i| D |w_j\rangle$ is zero for all $0 \leq i, j \leq n-2$, establishing in fact that $D = 0$. □

We use a slightly different convention for indexing the projectors, which is more natural, in the following.

§ 5.4 m Unitary | Solution to Mochon's m-assignments

There are four cases which arise. One could find a single expression for all of these but it does not seem to aid clarity. The four cases arise because there are two important parameters, the number of points and the degree of the monomial. These can be individually odd or even. We first consider the balanced case, where the number of points are even.

5.4.1 The Balanced Case

Even monomials (as opposed to odd monomials) seem to align well (at the bottom; see Figure 5.1a) and we start with them. The additional technique we introduce here is the use of $X_h^{\dashv}$ and $X_g^{\dashv}$ but the idea is essentially unchanged.

Proposition 87 (Solution to the balanced, even (aligned) monomial problem)**.** *Let*

- $m = 2b$ *be an even non-negative integer*
- *an* m*-assignment over* $0 < x_1 < x_2 \cdots < x_{2n}$ *be given by*

$$t = \sum_{i=1}^{n} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!],$$

- $\{|h_1\rangle, |h_2\rangle \dots |h_n\rangle, |g_1\rangle, |g_2\rangle \dots |g_n\rangle\}$ *be an orthonormal basis*
- *finally,*

$$X_h := \sum_{i=1}^{n} x_{h_i} |h_i\rangle \langle h_i| \doteq diag(x_{h_1}, x_{h_2} \dots x_{h_n}, \underbrace{0, 0 \dots 0}_{n \text{ zeros}}),$$

$$X_g := \sum_{i=1}^{n} x_{g_i} |g_i\rangle \langle g_i| \doteq diag(\underbrace{0, 0, \dots 0}_{n \text{ zeros}}, x_{g_1}, x_{g_2} \dots x_{g_n}),$$

$$|w\rangle := \sum_{i=1}^{n} \sqrt{p_{h_i}} |h_i\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots \sqrt{p_{h_n}}, \underbrace{0, 0, \dots 0}_{n \text{ zeros}})^T,$$

$$|v\rangle := \sum_{i=1}^{n} \sqrt{p_{g_i}} |g_i\rangle \doteq (\underbrace{0, 0, \dots 0}_{n \text{ zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots \sqrt{p_{g_n}})^T,$$

$|w'\rangle := (X_h)^b |w\rangle$ *and* $|v'\rangle := (X_g)^b |v\rangle$.

Then

$$O := \sum_{i=-b}^{n-b-1} \left(\frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle \langle v'| (X_g)^i \Pi_{g_i}^{\perp}}{\sqrt{c_{h_i} c_{g_i}}} + h.c. \right)$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad and \quad E_h O |v'\rangle = |w'\rangle$$

where for brevity, we used $(X_h)^{-k}$ *to mean* $(X_h^{\dashv})^k$ *(for* $k > 0$*),*

$$E_h := \sum_{i=1}^{n} |h_i\rangle \langle h_i|,$$

$$\Pi_{h_i}^{\perp} := \begin{cases} \text{projector orthogonal to span}\{(X_h)^{-|i|+1} |w'\rangle, (X_h)^{-|i|+2} |w'\rangle \dots, |w'\rangle\} & i < 0 \\ \text{projector orthogonal to span}\{(X_h)^{-b} |w'\rangle, (X_h)^{-b+1} |w'\rangle, \dots (X_h)^{i-1} |w'\rangle\} & i > 0 \\ \mathbb{I} & i = 0, \end{cases}$$

$c_{h_i} := \langle w'| (X_h)^i \Pi_{h_i}^{\perp} (X_h)^i |w'\rangle$ *and analogously*

$$\Pi_{g_i}^{\perp} := \begin{cases} \text{projector orthogonal to span}\{(X_g)^{-|i|+1} |v'\rangle, (X_g)^{-|i|+2} |v'\rangle \dots, |v'\rangle\} & i < 0 \\ \text{projector orthogonal to span}\{(X_g)^{-b} |v'\rangle, (X_g)^{-b+1} |v'\rangle, \dots (X_g)^{i-1} |v'\rangle\} & i > 0 \\ \mathbb{I} & i = 0, \end{cases}$$

$c_{g_i} := \langle v'| (X_g)^i \Pi_{g_i}^{\perp} (X_g)^i |v'\rangle$.

Proof. The orthonormal basis (over $\mathrm{span}\{|h_1\rangle, |h_2\rangle \ldots |h_n\rangle\}$) of interest here is

$$\left|w_i'\right\rangle := \frac{\Pi_{h_i}^{\perp}(X_h)^i \left|w'\right\rangle}{\sqrt{c_{h_i}}} \tag{5.7}$$

which entails

$$\Pi_{h_i}^{\perp} = \begin{cases} \mathbb{I}_h & i = 0 \\ \mathbb{I}_h - \sum_{j=i+1}^{0} \left|w_j'\right\rangle \left\langle w_j'\right| & i < 0 \\ \mathbb{I}_h - \sum_{j=-b}^{i-1} \left|w_j'\right\rangle \left\langle w_j'\right| & i > 0 \end{cases} \tag{5.8}$$

where $\mathbb{I}_h := E_h$. We define $|v_i'\rangle$ and $\Pi_{g_i}^{\perp}$ analogously. Our strategy would be to keep track of both the highest and lowest power, l in $\langle w'| X_h^l |w'\rangle$ and $\langle v'| X_g^l |v'\rangle$, which appear in the matrix elements $\langle w_i'| D \left|w_j'\right\rangle$. We use $\left\langle x_h^l \right\rangle' := \langle w'| X_h^l |w'\rangle = \langle w| X_h^{l+2b} |w\rangle$ and similarly $\left\langle x_g^l \right\rangle' := \langle v'| X_g^l |v'\rangle = \langle w| X_g^{l+2b} |w\rangle$. To this end, we denote the minimum and maximum powers, l, by

$$\mathcal{M}(\left|w_i'\right\rangle) = \begin{cases} \left(\left\langle x_h^0 \right\rangle' |w'\rangle, \left\langle x_h^0 \right\rangle' |w'\rangle\right) & i = 0 \\ \left(\left\langle x_h^{-2|i|} \right\rangle' (X_h)^{-|i|} |w'\rangle, \left\langle x_h^0 \right\rangle' |w'\rangle\right) & i < 0 \\ \left(\left\langle x_h^{-2b} \right\rangle' (X_h)^{-b} |w'\rangle, \left\langle x_h^{2i} \right\rangle' (X_h)^i |w'\rangle\right) & i > 0. \end{cases}$$

We define $D := X_h - E_h O X_g O^T E_h \doteq \langle w_i'| \left(X_h - E_h O X_g O^T E_h\right) \left|w_j'\right\rangle$. It suffices to restrict to the span of $\{|w_i'\rangle\}$ basis because $X_h |v_i'\rangle = 0$ and $E_h |v_i'\rangle = 0$. The lowest power, l, appearing in D is for $i = j = -b$ (as $-b \leq i, j \leq n - b - 1$). This can be evaluated to be $-2b$ by observing that

$$\mathcal{M}(\langle w_{-b}'|) X_h \mathcal{M}(\left|w_{-b}'\right\rangle) = \left(\left\langle x_h^{-2b} \right\rangle' \left\langle x_h^{-2b} \right\rangle' \left\langle x_h^{-2b+1} \right\rangle', \left\langle x_h^0 \right\rangle' \left\langle x_h^0 \right\rangle' \langle x_h\rangle'\right)$$

where we multiplied component-wise. To find the highest power, l, in the matrix D, note that for $i, j > 0$ we have

$$\mathcal{M}(\langle w_i'|) X_h \mathcal{M}(\left|w_j'\right\rangle) = \left(\left\langle x_h^{-2b} \right\rangle' \left\langle x_h^{-2b+1} \right\rangle' \left\langle x_h^{-2b} \right\rangle', \left\langle x_h^{2i} \right\rangle' \left\langle x_h^{2j} \right\rangle' \left\langle x_h^{i+j+1} \right\rangle'\right)$$

so $l = \max\{2i, 2j, i + j + 1\}$. As argued for the f_0-assignment, $l = 2n - 2b - 1$ for $i = j = n - b - 1$ and otherwise strictly less than $2n - 2b - 1$. Thus only the $D_{n-b-1,n-b-1}$ term in D, depends on $\left\langle x_h^{2n-2b-1} \right\rangle'$. Except for this term, all other terms, at most, depend on $\left\langle x_h^{-2b} \right\rangle', \left\langle x_h^{-2b+1} \right\rangle', \ldots \left\langle x_h^{2n-2b-2} \right\rangle'$, i.e. $\left\langle x_h^0 \right\rangle, \left\langle x_h^1 \right\rangle, \ldots \left\langle x_h^{2n-2} \right\rangle$. The analogous argument for $\langle v_i'| X_g \left|v_j'\right\rangle$, the observation that $\langle w_i'| D \left|w_j'\right\rangle = \langle w_i'| X_h \left|w_j'\right\rangle - \langle v_i'| X_g \left|v_j'\right\rangle$, and the fact that $\left\langle x^0 \right\rangle = \left\langle x^1 \right\rangle = \cdots = \left\langle x^{2n-2} \right\rangle = 0$ entail that these terms vanish. It remains to establish that $D_{n-b-1,n-b-1} \geq 0$. This is easily seen by noting that in $\langle w_{n-b-1}'| D \left|w_{n-b-1}'\right\rangle$, the only term which would not get cancelled due to the aforesaid reasoning, must come from the part of $\left|w_{n-b-1}'\right\rangle$ containing $X_h^{n-b-1} |w'\rangle$. It suffices to show that the coefficient of this term is positive because we know that $\left\langle x^{2n-2b-1} \right\rangle' = \left\langle x^{2n-1} \right\rangle > 0$. Further, we know the coefficient exactly: it is $1/c_{h_{n-b-1}}$ (see Equation (5.8) and Equation (5.7)). This establishes that $D \geq 0$. □

To proceed further, it is helpful to have a more concise way of viewing the proof. To this end, we consider a concrete example of a balanced (aligned) m-assignment, with $2n = 8$ and $m = 2b = 2$ (see Figure 5.1a). We represent the range of dependence of $\langle w_0'| X_h |w_0'\rangle$ on $\left\langle x_h^l \right\rangle$ diagrammatically by

enclosing in a left bracket, the terms $\langle x^3\rangle = \langle x\rangle'$ and $\langle x^2\rangle = \langle x^0\rangle'$ (replacing $|w\rangle$ with $|w_0'\rangle$) and writing $|w_0'\rangle$ next to it. Similarly, for $|w_{-1}'\rangle, |w_1'\rangle$ and $|w_2'\rangle$ we enclose in a left bracket, the terms

$$\{\langle x^0\rangle, \langle x^1\rangle, \langle x^2\rangle, \langle x^3\rangle\} = \left\{\langle x^{-2}\rangle', \langle x^{-1}\rangle', \ldots \langle x\rangle'\right\},$$

$$\{\langle x^0\rangle, \langle x^1\rangle, \ldots, \langle x^5\rangle\} = \left\{\langle x^{-2}\rangle', \langle x^{-1}\rangle', \ldots \langle x^3\rangle'\right\}$$

and

$$\{\langle x^0\rangle, \langle x^1\rangle, \ldots \langle x^7\rangle\} = \left\{\langle x^{-2}\rangle', \langle x^{-1}\rangle', \ldots \langle x^5\rangle'\right\}$$

respectively. Keep in mind that the highest power l of $\langle x_h^l\rangle$ that appears in $\langle w_i'| X_h \left|w_j'\right\rangle$ is $l = 7$ when (and only when) $i = j = 2$. Thus, the matrix D, restricted to the subspace spanned by the $\{|w_i'\rangle\}$ basis (again, we can safely ignore the subspace span$\{|v_i'\rangle\}$ because $D|v_i'\rangle = 0$), has only one non-zero entry which we saw was positive as $\langle x^7\rangle > 0$. This is reminiscent of the balanced f_0-assignment (and includes it as a special case).

We now explain why a direct extension of the analysis to the balanced (misaligned) m-assignment fails and subsequently see how to remedy the situation. Consider, for concreteness, the case with $2n = 8$ and $m = 2b - 1 = 3$ (see Figure 5.1b). From hindsight, we write both the $|v_i'\rangle$s and the $|w_i'\rangle$s. We start with $|w_0'\rangle = X_h^{3/2}|w\rangle$ and $|v_0'\rangle = X_g^{3/2}|v_0\rangle$, and as before, enclose the terms $\left\{\langle x^0\rangle' = \langle x^3\rangle, \langle x^1\rangle' = \langle x^4\rangle\right\}$ in a left bracket. We then go below by multiplying $|w_0'\rangle$ with X_h^{-1} (and $|v_0'\rangle$ with X_g^{-1} respectively) and projecting out the components along the previous vectors. We represent these by $\left|w_{-1}'\right\rangle$ and $\left|v_{-1}'\right\rangle$ and in the figure, enclose the terms $\left\{\langle x\rangle = \langle x^{-2}\rangle', \langle x^2\rangle = \langle x^{-1}\rangle' \ldots \langle x^4\rangle = \langle x\rangle'\right\}$ in the left and right brackets. We do not go lower because then we pickup a dependence on $\langle x^{-1}\rangle$ which will persist for subsequent vectors. In general, we stop after taking b (which equals 1 here) steps down. We go up by multiplying with $|w_0'\rangle$ with X_h (and $|v_0'\rangle$ with X_g resp.) and projecting out the components along the previous vectors. We represent these by $|w_1'\rangle$ and in the figure, $|v_1'\rangle$ and enclose the terms $\left\{\langle x\rangle = \langle x^{-2}\rangle', \langle x^2\rangle = \langle x^{-1}\rangle' \ldots \langle x^6\rangle = \langle x^3\rangle'\right\}$ in the brackets. Finally, we construct $|w_2'\rangle$ and $|v_2'\rangle$ by taking a step up using X_h and X_g resp. (these are essentially fixed to be the vectors orthogonal to the previous ones once we restrict to span$(|h_1\rangle, |h_2\rangle \ldots |h_n\rangle)$) and span$(|g_1\rangle, |g_2\rangle \ldots |g_n\rangle)$). Taking a step down using X_h^{-1} and X_g^{-1} we could have constructed $\left|w_{-2}'\right\rangle$ and $\left|v_{-2}'\right\rangle$ respectively but they are the same as $|w_2'\rangle$ and $|v_2'\rangle$ respectively (as explained). If we were to use $O = \sum_{i=-1}^{2}(|w_i'\rangle\langle v_i'| + \text{h.c.})$ then we would have obtained dependence on $\langle x^7\rangle$ in the last row (corresponding to $|w_2'\rangle$) and a dependence on $\langle x^8\rangle$ for the last term (i.e. $\langle w_2'|D|w_2'\rangle$). This already hints that the matrix is negative because it has the form $\begin{bmatrix} 0 & b \\ b & c \end{bmatrix}$ with $b \neq 0$ which means that the determinant is $-b^2$, entailing there's a negative eigenvalue; thus this choice can not work. We therefore define $O := \left(\sum_{i=-1}^{1}|w_i'\rangle\langle v_i'| + \text{h.c.}\right) + |w_2'\rangle\langle w_2'| + |v_2'\rangle\langle v_2'|$ which acts as a 'blinkered unitary' on the remaining troublesome parts. Further, instead of using

$$X_h \geq E_h O X_g O^T E_h \tag{5.9}$$

for establishing positivity, we equivalently use

$$E_h \geq \left(X_h^{\dashv}\right)^{1/2} O X_g O^T \left(X_h^{\dashv}\right)^{1/2}, \tag{5.10}$$

which is easily obtained by multiplying by $(X_h^{\dashv})^{1/2}$ from both the sides. The reason is that to establish positivity, we must include $|w_2'\rangle$ in the basis (we can neglect the null vectors of E_h), and even though the RHS (of Equation (5.9)) would not contribute, the LHS would get non-trivial contributions along the

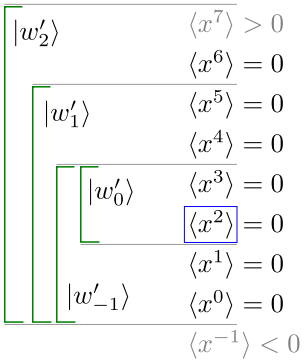
rows (as was the case earlier). Using the form with the inverses lets us remove this dependence. To see this, note that span$\{\left|w'_{-1}\right\rangle, |w'_0\rangle \dots |w'_2\rangle\}$ equals the h-space, i.e. span$\{|h_1\rangle, |h_2\rangle \dots |h_n\rangle\}$. Further, span$\{X_h^{1/2} |w'_i\rangle\}_{i=-1}^{2}$ also equals the h-space (but the vectors are not, in general, orthonormal any more). Finally, observe that $X_h^{1/2} |w'_2\rangle$ is a null vector of the RHS of Equation (5.10). These entail that to prove the positivity, it suffices to restrict to span$\{X_h^{1/2} |w'_i\rangle\}_{i=-1}^{1}$. An arbitrary normalised vector in this space can be written as

$$|\psi\rangle = \frac{\sum_{i=-1}^{1} \alpha_i X_h^{1/2} |w'_i\rangle}{\sqrt{\sum_{i,j=-1}^{1} \alpha_i \alpha_j \langle w'_i| X_h \left|w'_j\right\rangle}}$$

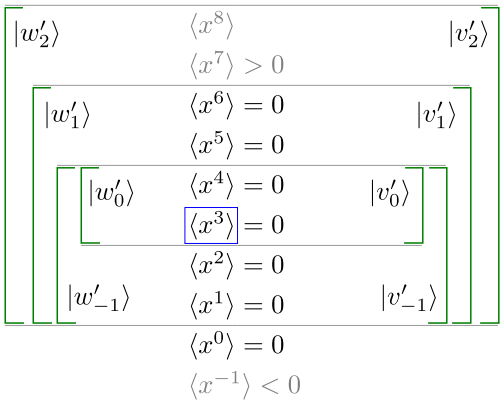
$$\implies X_g^{1/2} O^T (X_h^{\dashv})^{1/2} |\psi\rangle = \frac{\sum_{i=-1}^{1} \alpha_i X_g^{1/2} |v'_i\rangle}{\sqrt{\sum_{i,j=-1}^{1} \alpha_i \alpha_j \langle w'_i| X_h \left|w'_j\right\rangle}}$$

$$\implies \langle\psi| (X_h^{\dashv})^{1/2} O X_g O^T (X_h^{\dashv})^{1/2} |\psi\rangle = \frac{\sum_{i,j=-1}^{1} \alpha_i \alpha_j \langle v'_i| X_g \left|v'_j\right\rangle}{\sum_{i,j=-1}^{1} \alpha_i \alpha_j \langle w'_i| X_h \left|w'_j\right\rangle} = 1.$$

where we get equality by noting that $\langle v'_i| X_g \left|v'_j\right\rangle$s depend on (at most) $\{\langle x_g\rangle, \langle x_g^2\rangle \dots \langle x_g^6\rangle\}$ and analogously $\langle w'_i| X_h \left|w'_j\right\rangle$ depend on (at most) $\{\langle x_h\rangle, \langle x_h^2\rangle \dots \langle x_h^6\rangle\}$, we conclude they are the same as $\langle x^i\rangle = 0$ for $i \in \{0, 1, \dots 6\}$. Since we proved the RHS (of Equation (5.10)) is one for all normalised $|\psi\rangle$s, we conclude we have the correct unitary. Note that this technique can also be used for the previous results. We now give the general statement and proof.



(a) $2n = 8$, $m = 2b = 2$; Even (aligned) balanced m-assignment.



(b) $2n = 8$, $m = 2b - 1 = 3$; Odd (misaligned) balanced m-assignment.

Figure 5.1: Visualising balanced monomial assignments with simple examples.

Proposition 88 (Solution to the balanced, odd (misaligned) monomial problem)**.** *Let*

- $m = 2b - 1$ *be an odd non-negative integer (i.e.* $b \geq 1$*)*
- *an* m*-assignment over* $\{x_1, x_2 \dots x_{2n}\}$ *be given by*

$$t = \sum_{i=1}^{n} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!],$$

- $(|h_1\rangle, |h_2\rangle \dots |h_n\rangle, |g_1\rangle, |g_2\rangle \dots |g_n\rangle)$ *be an orthonormal basis*

- *finally*

$$X_h := \sum_{i=1}^{n} x_{h_i} |h_i\rangle \langle h_i| \doteq \mathit{diag}(x_{h_1}, x_{h_2} \dots x_{h_n}, \underbrace{0, 0 \dots 0}_{n \text{ zeros}}),$$

$$X_g := \sum_{i=1}^{n} x_{g_i} |g_i\rangle \langle g_i| \doteq \mathit{diag}(\underbrace{0, 0, \dots 0}_{n \text{ zeros}}, x_{g_1}, x_{g_2} \dots x_{g_n}),$$

$$|w\rangle := (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots \sqrt{p_{h_n}}, \underbrace{0, 0 \dots 0}_{n \text{ zeros}}),$$

$$|v\rangle := (\underbrace{0, 0, \dots 0}_{n \text{ zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots \sqrt{p_{g_n}}),$$

$|w'\rangle := (X_h)^{b-\frac{1}{2}} |w\rangle$ *and* $|v'\rangle := (X_g)^{b-\frac{1}{2}} |v\rangle$.

Then

$$O := \sum_{i=-b+1}^{n-b-1} \left(\frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle \langle v'| (X_g)^i \Pi_{g_i}^{\perp}}{\sqrt{c_{h_i} c_{g_i}}} + h.c. \right) + \frac{\Pi_{g_{n-b}}^{\perp} (X_g)^{n-b} |v'\rangle \langle v'| (X_g)^{n-b} \Pi_{g_{n-b}}^{\perp}}{c_{g_{n-b+1}}} + \frac{\Pi_{h_{n-b}}^{\perp} (X_h)^{n-b} |w'\rangle \langle w'| (X_h)^{n-b} \Pi_{h_{n-b}}^{\perp}}{c_{h_{n-b}}}$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad \textit{and} \quad E_h O |v'\rangle = |w'\rangle$$

where for brevity, by X_h^{-k} *we mean* $(X_h^{\dashv})^k$ *for* $k > 0$ *(similarly for* X_g*),* $c_{h_i} := \langle w'| (X_h)^i \Pi_{h_i}^{\perp} (X_h)^i |w'\rangle$, $c_{g_i} := \langle v'| (X_g)^i \Pi_{g_i}^{\perp} (X_g)^i |v'\rangle$,

$$\Pi_{h_i}^{\perp} := \begin{cases} \textit{projector orthogonal to span}\{(X_h^{\dashv})^{|i|-1} |w'\rangle, (X_h^{\dashv})^{|i|-2} |w'\rangle \dots, |w'\rangle\} & i < 0 \\ \textit{projector orthogonal to span}\{(X_h^{\dashv})^{b-1} |w'\rangle, (X_h^{\dashv})^{b-2} |w'\rangle, \dots, |w'\rangle, X_h |w'\rangle, \dots (X_h)^{i-1} |w'\rangle\} & i > 0 \\ \mathbb{I} & i = 0, \end{cases}$$

and analogously

$$\Pi_{g_i}^{\perp} := \begin{cases} \textit{projector orthogonal to span}\{(X_g^{\dashv})^{|i|-1} |v'\rangle, (X_g^{\dashv})^{|i|-2} |v'\rangle \dots, |v'\rangle\} & i < 0 \\ \textit{projector orthogonal to span}\{(X_g^{\dashv})^{b-1} |v'\rangle, (X_g^{\dashv})^{b-2} |v'\rangle, \dots |v'\rangle, X_g |v'\rangle, \dots (X_g)^{i-1} |v'\rangle\} & i > 0 \\ \mathbb{I} & i = 0. \end{cases}$$

Proof. The proof is very similar to that of Proposition 87. For brevity, in this proof as well, by X_h^{-k} we mean $(X_h^{\dashv})^k$ for $k > 0$ (similarly for X_g). The orthonormal basis (over $\{|h_1\rangle, |h_2\rangle \dots |h_n\rangle\}$) of interest here is

$$|w_i'\rangle := \frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle}{\sqrt{c_{h_i}}}$$

which entails

$$\Pi_{h_i}^{\perp} = \begin{cases} \mathbb{I}_h & i = 0 \\ \mathbb{I}_h - \sum_{j=i-1}^{0} \left|w_j'\right\rangle \left\langle w_j'\right| & i < 0 \\ \mathbb{I}_h - \sum_{j=-b+1}^{i} \left|w_j'\right\rangle \left\langle w_j'\right| & i > 0 \end{cases}$$

where $\mathbb{I}_h := E_h$. We define $|v_i'\rangle$ and $\Pi_{g_i}^{\perp}$ analogously. Our strategy is to keep track of the highest and lowest powers, l in $\langle w'| X_h^l |w'\rangle$ and $\langle v'| X_g^l |v'\rangle$, which appear in the matrix elements $\langle w_i'| X_h \left|w_j'\right\rangle$ and $\langle v_i'| X_g \left|v_j'\right\rangle$. For brevity, as before, we use $\left\langle x_h^l \right\rangle' := \langle w'| X_h^l |w'\rangle$ and similarly $\left\langle x_g^l \right\rangle' := \langle v'| X_g^l |v'\rangle$. To this end, we denote the minimum and maximum powers, l, by

$$\mathcal{M}(|w_i'\rangle) = \begin{cases} \left(\left\langle x_h^0 \right\rangle' |w'\rangle, \left\langle x_h^0 \right\rangle' |w'\rangle\right) & i = 0 \\ \left(\left\langle x_h^{-2|i|} \right\rangle' (X_h)^{-|i|} |w'\rangle, \left\langle x_h^0 \right\rangle' |w'\rangle\right) & i < 0 \\ \left(\left\langle x_h^{-2(b-1)} \right\rangle' (X_h)^{-(b-1)} |w'\rangle, \left\langle x_h^{2i} \right\rangle' (X_h)^{i} |w'\rangle\right) & i > 0. \end{cases}$$

Note that establishing $X_h \geq E_h O X_g O^T E_h$ is equivalent to establishing

$$E_h \geq X_h^{-1/2} O X_g O^T X_h^{-1/2} \tag{5.11}$$

where, note that by $X_h^{-1/2}$ we mean $(X_h^{-1})^{1/2}$. It is easy to see that $X_h^{1/2} \left|w_{n-b}'\right\rangle$ is a null vector (vector with zero eigenvalue) for the RHS as $X_g O^T \left|w_{n-b}'\right\rangle = 0$. Any vector, $|\psi\rangle$ in $\text{span}\{|g_1\rangle, |g_2\rangle \ldots |g_n\rangle\}$ is a null vector for both the LHS and the RHS. Thus, to establish the positivity, we can restrict to $\text{span}\{|h_1\rangle, |h_2\rangle, \ldots |h_n\rangle\} \backslash \text{span}\{X_h^{1/2} \left|w_{n-b}'\right\rangle\}$, i.e. to vectors in the h-space orthogonal to $X_h^{1/2} \left|w_{n-b}'\right\rangle$. It turns out to be easier to test for positivity on a possibly larger space. It is clear that $\text{span}\left\{X_h^{1/2} |w_i'\rangle\right\}_{i=-b+1}^{n-b}$ equals $\text{span}\{|h_1\rangle, |h_2\rangle \ldots |h_n\rangle\}$ (because it also equals $\text{span}\{|w'\rangle_i\}_{i=-b+1}^{n-b}$, due to Lemma 85). As neglecting vectors with components along $X_h^{1/2} \left|w_{n-b}'\right\rangle$ suffices (for establishing positivity of Equation (5.11)), we can restrict to $\text{span}\{X_h^{1/2} |w_i'\rangle\}_{i=-b+1}^{n-b-1}$ (which might still contain vectors with components along $X_h^{1/2} \left|w_{n-b}'\right\rangle$ as the basis vectors are not orthogonal but it only means that we check for positivity over a larger set of vectors). These ensure that the troublesome vectors (neither $\left|w_{n-b}'\right\rangle$ nor $\left|v_{n-b}'\right\rangle$), appear in the remaining analysis. Let $|\psi\rangle = \left(\sum_{i=-b+1}^{n-b-1} \alpha_i X_h^{1/2} |w_i'\rangle\right)/c$ where $c = \sqrt{\langle\psi|\psi\rangle}$. To establish Equation (5.11), it is enough to show that for all choices of α_is,

$$\begin{aligned} 1 &\geq \langle\psi| X_h^{-1/2} O X_g O^T X_h^{-1/2} |\psi\rangle \\ &= \frac{\sum_{i,j=-b+1}^{n-b-1} \alpha_i \alpha_j \langle v_i'| X_g \left|v_j'\right\rangle}{\sum_{i,j=-b+1}^{n-b-1} \alpha_i \alpha_j \langle w_i'| X_h \left|w_j'\right\rangle} \\ &= 1 \end{aligned} \tag{5.12}$$

where the second step follows from noting that $X_g^{1/2} O^T X_h^{-1/2} |\psi\rangle = \sum_{i=-b+1}^{n-b-1} \alpha_i X_g^{1/2} |v_i'\rangle$ and the last step follows from a counting argument which we now give.

Note that

$$\left\langle x_h^i \right\rangle' = \left\langle x_h^{i+2b-1} \right\rangle \tag{5.13}$$

and that

$$\left\langle x^0 \right\rangle = \langle x \rangle = \cdots = \left\langle x^{2n-2} \right\rangle = 0. \tag{5.14}$$

To determine the highest power of l in $\langle w'| X_h^l |w'\rangle$ which appears in the matrix elements $\langle w_i'| X_h \left|w_j'\right\rangle$ (for $-b+1 \leq i, j \leq n-b-1$) it suffices to consider $\left\langle w_{n-b-1}'\right| X_h \left|w_{n-b-1}'\right\rangle$. To this end, we evaluate

$$\begin{aligned} &\mathcal{M}(\left\langle w_{n-b-1}'\right|) X_h \mathcal{M}(\left|w_{n-b-1}'\right\rangle) \\ &= \left(\left\langle x_h^{-2(b-1)} \right\rangle' \left\langle x_h^{-2(b-1)} \right\rangle' \left\langle x_h^{-2(b-1)+1} \right\rangle', \left\langle x_h^{2(n-b-1)} \right\rangle' \left\langle x_h^{2(n-b-1)} \right\rangle' \left\langle x_h^{2(n-b-1)+1} \right\rangle'\right) \\ &= \left(\langle x_h \rangle \langle x_h \rangle \left\langle x_h^2 \right\rangle, \left\langle x_h^{2n-3} \right\rangle \left\langle x_h^{2n-3} \right\rangle \left\langle x_h^{2n-2} \right\rangle\right). \end{aligned}$$

The highest power is, manifestly, $l = 2n-2$. To find the lowest power l in $\langle w'| X_h^l |w'\rangle$ which appears in the matrix elements $\langle w'_i| X_h \left|w'_j\right\rangle$ (for $-b+1 \leq i,j \leq n-b-1$) it suffices to consider $\left\langle w'_{-b+1}\right| X_h \left|w'_{-b+1}\right\rangle$. To this end, we evaluate

$$\begin{aligned}\mathcal{M}(\left\langle w'_{-b+1}\right|)X_h\mathcal{M}(\left|w'_{-b+1}\right\rangle) &= \left(\left\langle x_h^{-2(b-1)}\right\rangle' \left\langle x_h^{-2(b-1)}\right\rangle' \left\langle x_h^{-2(b-1)+1}\right\rangle', \langle x_h^0\rangle' \langle x_h^0\rangle' \langle x_h\rangle'\right) \\ &= \left(\langle x_h\rangle \langle x_h\rangle \left\langle x_h^2\right\rangle, \left\langle x_h^{2b-1}\right\rangle \left\langle x_h^{2b-1}\right\rangle \left\langle x_h^{2b}\right\rangle\right).\end{aligned}$$

The lowest power is, manifestly, $l = 1$. We thus conclude that the numerator of Equation (5.12) is a function of $\langle x_h\rangle, \left\langle x_h^2\right\rangle, \ldots \left\langle x_h^{2n-2}\right\rangle$and, an analogous argument entails that the denominator is a function of $\langle x_g\rangle, \left\langle x_g^2\right\rangle, \ldots \left\langle x_g^{2n-2}\right\rangle$ with the same form. Using Equation (5.14), we conclude that they (the numerator and denominator) are the same. □

5.4.2 The Unbalanced Case

The techniques we have used so far also work when the number of points in an m-assignment are odd, i.e. unbalanced m-assignments. We distinguish unbalanced m-assignments as odd (misaligned) or even (aligned) depending on the power of the monomial. We illustrate how the solution is constructed by considering a concrete example of an unbalanced even (aligned) m-assignment. We start with $2n-1 = 7$ points and $m = 2b = 2$ (see Figure 5.2a). We use the diagrammatic representation introduced in the previous (sub)section (after Proposition 87). In this case, we have 4 initial points and 3 final points; the standard basis is $\{|g_1\rangle, |g_2\rangle, \ldots |g_4\rangle, |h_1\rangle, |h_2\rangle, |h_3\rangle\}$. The basis of interest is, as usual, constructed by starting at $|w'\rangle$ (and analogously for $|v'\rangle$) and using X_h^{-1} first until we reach $\langle x^0\rangle$ followed by using X_h until the space is spanned. It is $\left\{\left|v'_{-1}\right\rangle, |v'_0\rangle, |v'_1\rangle, |v'_2\rangle\right\}$ and $\left\{\left|w'_{-1}\right\rangle, |w'_0\rangle, |w'_1\rangle\right\}$. In the same vein as the earlier solutions, we define $O := \sum_{i=-1}^{1} (|w'_i\rangle \langle v'_i| + \text{h.c.}) + |v'_2\rangle \langle v'_2|$. In $X_h \geq E_h O X_g O^T E_h$, the $|v'_2\rangle$ term is removed by the projector, $E_h := \sum_{i=1}^{3} |h_i\rangle \langle h_i|$. Using $\langle x^0\rangle = \langle x\rangle = \cdots = \langle x^5\rangle = 0$ and the counting arguments from before, it follows that $D = X_h - E_h O X_g O^T E_h$ is zero.

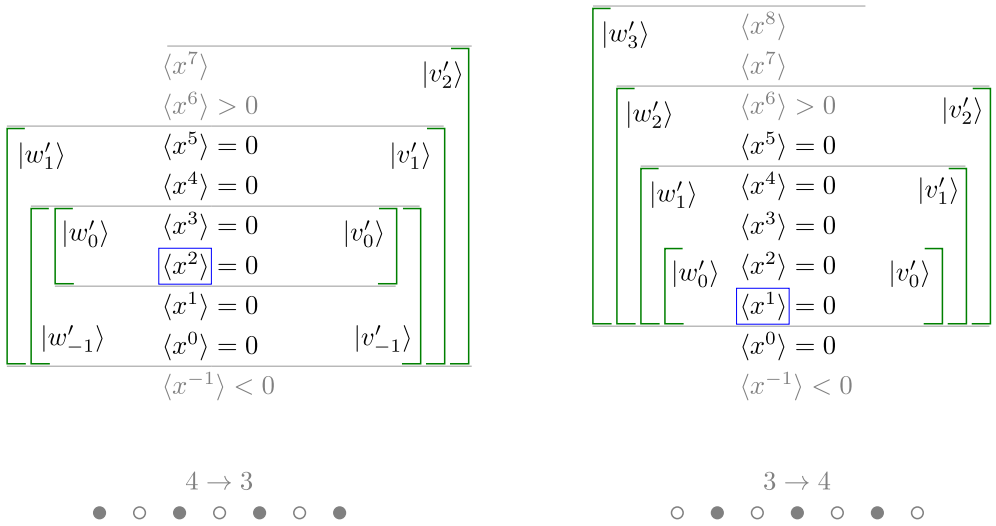
We now illustrate how the solution is constructed by considering an unbalanced odd m-assignment. Consider a concrete example with $2n-1 = 7$ and $m = 2b-1 = 1$. In this case, we have 3 initial points and 4 final points; the standard basis is $\{|g_1\rangle, |g_2\rangle, |g_3\rangle, |h_1\rangle, |h_2\rangle, \ldots |h_4\rangle\}$. The basis of interest is, in this case, constructed by starting at $|w'\rangle$ (and analogously for $|v'\rangle$) and using X_h until the space is spanned. More generally, we first go below for $b-2$ steps (which is zero in this case), until $\langle x\rangle$, is reached in the diagram. The basis is $\{|v'_0\rangle, |v'_1\rangle, |v'_2\rangle\}$ and $\{|w'_0\rangle, |w'_1\rangle, |w'_2\rangle, |w'_3\rangle\}$. As before, we define $O := \sum_{i=0}^{2} (|w'_i\rangle \langle v'_i| + \text{h.c.}) + |w'_3\rangle \langle w'_3|$. This time we[2] use $E_h \geq X_h^{-1/2} O X_g O^T X_h^{-1/2}$ which is equivalent to $X_h \geq E_h O X_g O^T E_h$ for $E_h := \sum_{i=1}^{4} |h_i\rangle \langle h_i|$. Using an argument similar to the balanced misaligned case, we can reduce the positivity condition to

$$1 \geq \frac{\sum_{i,j=0}^{2} \alpha_i \alpha_j \langle v'_i| X_g \left|v'_j\right\rangle}{\sum_{i,j=0}^{2} \alpha_i \alpha_j \langle w'_i| X_h \left|w'_j\right\rangle}$$

but the counting argument doesn't make the fraction 1. This is because we now have an $\left\langle x_h^6\right\rangle$ dependence in the denominator (and $\left\langle x_g^6\right\rangle$ dependence in the numerator). However, we also know that this term only appears in $\langle w'_2| X_h |w'_2\rangle$ that too with a positive coefficient (as we saw in the unbalanced f_0 assignment). Further, we know $\left\langle x_h^6\right\rangle > \left\langle x_g^6\right\rangle$ and therefore we can conclude that the numerator is smaller than the denominator ensuring the inequality is always satisfied.

We state the general solution for both these cases and prove their correctness below.

[2] remember by $X_h^{-1/2}$ we mean $(X_h^{\dashv})^{1/2}$



(a) $2n-1=7$; $m=2b=2$. Even (aligned) unbalanced m-assignment.

(b) $2n-1=7$; $m=2b-1=1$. Odd (misaligned) unbalanced m-assignment.

Figure 5.2: Visualising unbalanced m-assignment with simple examples.

Proposition 89 (Solution to the unbalanced, even (aligned) monomial problem)**.** *Let*

- $m = 2b$ *be an even non-negative integer*
- *an* m*-assignment over* $\{x_1, x_2 \dots x_{2n-1}\}$ *be given by*

$$t = \sum_{i=1}^{n-1} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!],$$

- $(|h_1\rangle, |h_2\rangle \dots |h_{n-1}\rangle, |g_1\rangle, |g_2\rangle \dots |g_n\rangle)$ *be an orthonormal basis*
- *finally*

$$X_h := \sum_{i=1}^{n-1} x_{h_i} |h_i\rangle\langle h_i| \doteq \mathit{diag}(x_{h_1}, x_{h_2} \dots x_{h_{n-1}}, \underbrace{0, 0 \dots 0}_{n \text{ zeros}}),$$

$$X_g := \sum_{i=1}^{n} x_{g_i} |g_i\rangle\langle g_i| \doteq \mathit{diag}(\underbrace{0, 0, \dots 0}_{n-1 \text{ zeros}}, x_{g_1}, x_{g_2} \dots x_{g_n}),$$

$$|w\rangle := (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots \sqrt{p_{h_{n-1}}}, \underbrace{0, 0 \dots 0}_{n \text{ zeros}}),$$

$$|v\rangle := (\underbrace{0, 0, \dots 0}_{n-1 \text{ zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots \sqrt{p_{g_n}}),$$

$|w'\rangle := (X_h)^b |w\rangle$ *and* $|v'\rangle := (X_g)^b |v\rangle$.

Then

$$O := \sum_{i=-b}^{n-b-2} \left(\frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle\langle v'| (X_g)^i \Pi_{g_i}^{\perp}}{\sqrt{c_{h_i} c_{g_i}}} + \mathit{h.c.} \right) + \frac{\Pi_{g_{n-b-1}}^{\perp} (X_g)^{n-b-1} |v'\rangle\langle v'| (X_g)^{n-b-1} \Pi_{g_{n-b-1}}^{\perp}}{c_{g_{n-b-1}}}$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad \textit{and} \quad E_h O \left|v'\right\rangle = \left|w'\right\rangle$$

where for brevity, by X_h^{-k} we mean $(X_h^{\dashv})^k$ for $k > 0$ (similarly for X_g), $c_{h_i}, c_{g_i}, \Pi_{h_i}^{\perp}, \Pi_{g_i}^{\perp}$ are as defined in Proposition 87.

Proof. Many observations from the proof of Proposition 87 carry over to this case. We import the definitions of $\{|w_i'\rangle\}_{i=-b}^{n-b-2}$ and $\{|v_i'\rangle\}_{i=-b}^{n-b-1}$, together with the observations that $\mathcal{M}(\langle w'_{-b}|)X_h\mathcal{M}(|w'_{-b}\rangle)$ has no dependence on a term $\langle x_h^l\rangle'$ with l smaller than $-2b$ (which corresponds to $\langle x_h\rangle$) and that $\mathcal{M}(\langle w'_{n-b-2}|)X_h\mathcal{M}(|w'_{n-b-2}\rangle)$ has no dependence on a term $\langle x_h^l\rangle'$ with l greater than $2n-2b-4+1 = 2n-3-2b$. We can restrict to $\text{span}\{|w'_{-b}\rangle, |w'_{-b+1}\rangle \ldots |w'_{n-b-2}\rangle\}$ to establish the positivity of $D := X_h - E_h O X_g O^T E_h$. Using the analogous observation for $\mathcal{M}(\langle v'_{-b}|)X_g\mathcal{M}(|v'_{-b}\rangle)$ and $\mathcal{M}(\langle v'_{n-b-2}|)X_g\mathcal{M}(|v'_{n-b-2}\rangle)$, along with the fact that $\langle x^l\rangle' = \langle x^{l+2b}\rangle$, and $\langle x^0\rangle = \langle x^1\rangle = \cdots = \langle x^{2n-3}\rangle = 0$, it follows that D is zero. □

Proposition 90 (Solution to the unbalanced, odd (misaligned) monomial problem)**.** *Let*

- $m = 2b-1$ *be an odd non-negative integer*
- *an m-assignment over $\{x_1, x_2 \ldots x_{2n-1}\}$ be given by*

$$t = \sum_{i=1}^{n} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n-1} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!],$$

- $(|h_1\rangle, |h_2\rangle \ldots |h_n\rangle, |g_1\rangle, |g_2\rangle \ldots |g_{n-1}\rangle)$ *be an orthonormal basis*
- *finally*

$$X_h := \sum_{i=1}^{n} x_{h_i} |h_i\rangle\langle h_i| \doteq \textit{diag}(x_{h_1}, x_{h_2} \ldots x_{h_n}, \underbrace{0, 0 \ldots 0}_{n-1 \text{ zeros}}),$$

$$X_g := \sum_{i=1}^{n-1} x_{g_i} |g_i\rangle\langle g_i| \doteq \textit{diag}(\underbrace{0, 0, \ldots 0}_{n \text{ zeros}}, x_{g_1}, x_{g_2} \ldots x_{g_{n-1}}),$$

$$|w\rangle := (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \ldots \sqrt{p_{h_n}}, \underbrace{0, 0 \ldots 0}_{n-1 \text{ zeros}}),$$

$$|v\rangle := (\underbrace{0, 0, \ldots 0}_{n \text{ zeros}}, \sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_{n-1}}}),$$

$|w'\rangle := (X_h)^{b-\frac{1}{2}} |w\rangle$ *and* $|v'\rangle := (X_g)^{b-\frac{1}{2}} |v\rangle$.

Then

$$O := \sum_{i=-b+1}^{n-b-1} \left(\frac{\Pi_{h_i}^{\perp} (X_h)^i |w'\rangle\langle v'| (X_g)^i \Pi_{g_i}^{\perp}}{\sqrt{c_{h_i} c_{g_i}}} + h.c. \right) + \frac{\Pi_{h_{n-b}}^{\perp} (X_h)^{n-b} |w'\rangle\langle w'| (X_h)^{n-b} \Pi_{h_{n-b}}^{\perp}}{c_{h_{n-b}}}$$

satisfies

$$X_h \geq E_h O X_g O^T E_h \quad \textit{and} \quad E_h O \left|v'\right\rangle = \left|w'\right\rangle$$

where for brevity, by X_h^{-k} we mean $(X_h^{\dashv})^k$ for $k > 0$ (similarly for X_g), $c_{h_i}, c_{g_i}, \Pi_{h_i}^{\perp}, \Pi_{g_i}^{\perp}$ are as defined in Proposition 88.

Proof. For this proof, we can use the definitions and observations from the proof of Proposition 88. We import the definitions of $\{|w'_i\rangle\}_{i=-b+1}^{n-b}$ and $\{|v'_i\rangle\}_{i=-b+1}^{n-b-1}$ along with the observation that

$$\mathcal{M}(\langle w'_{-b+1}|)X_h\mathcal{M}(|w'_{-b+1}\rangle)$$

has no dependence on a term $\langle x_h^l\rangle'$ with l less than $-2b+2$ (which corresponds to $\langle x_h\rangle$) and

$$\mathcal{M}(\langle w'_{n-b-1}|)X_h\mathcal{M}(|w'_{n-b-1}\rangle)$$

has no dependence on a term $\langle x^l\rangle$ with l greater than $2n-2b-1$ (which corresponds to $\langle x_h^{2n-2}\rangle$ because $2n-2b-1+(2b-1)=2n-2$). We can also conclude, from the previous proof, that establishing $X_h \geq E_h O X_g O^T E_h$ is equivalent to establishing that

$$1 \geq \frac{\sum_{i,j=-b+1}^{n-b-1}\alpha_i\alpha_j\langle v'_i|X_g|v'_j\rangle}{\sum_{i,j=-b+1}^{n-b-1}\alpha_i\alpha_j\langle w'_i|X_h|w'_j\rangle}$$

for all (real) $\{\alpha_i\}_{i=-b+1}^{n-b-1}$. We know that $\langle x\rangle = \langle x^2\rangle = \cdots = \langle x^{2n-3}\rangle = 0$. As we have dependence on $\langle x_h^{2n-2}\rangle$, we can't conclude that the fraction is one. However, as we saw in the proof of Proposition 87, dependence on $\langle x_h^{2n-2}\rangle$ (in the denominator) only appears in the $\langle w'_{n-b-1}|X_h|w'_{n-b-1}\rangle$ term, that too with the positive coefficient, $1/c_{h_{n-b-1}}$. The analogous statement holds for the numerator. This, using $\langle x^{2n-2}\rangle > 0$, entails that the denominator is larger than or equal to the numerator, concluding the proof. □

§ 5.5 Main Result

Our observations so far can be combined to prove Theorem 11, which we formally state here.

Theorem. *Let t be Mochon's f-assignment (see Definition 80) on strictly positive coordinates without loss of generality (see Lemma 83). Suppose f has real and strictly positive roots. Then, decompose the assignment as $t = \sum_i \alpha_i t'_i$ where α_i are positive and t'_i are monomial assignments (see Lemma 82). Each t'_i admits an exact solution[3] given in Proposition 87, Proposition 88, Proposition 89, or Proposition 90, depending on t'_i.*

Proof. From Lemma 82, we deduce that t can be expressed as a sum of monomials. A monomial assignment can be categorised into four—balanced/unbalanced aligned/misaligned. The solution in each case is given by either Proposition 87, Proposition 88, Proposition 89, or Proposition 90. □

5.5.1 Example: A bias $1/14$ protocol

The $1/14$ bias unitaries

The typical move/assignment that Mochon's bias $\epsilon(3) = 1/14$ ($k=3$ for $\epsilon(k) = \frac{1}{4k+2}$) TIPG uses has the following form: Let

$$x'_0 = 0 < r'_1 < r'_2 < x'_1 < x'_2 < x'_3 < x'_4 < x'_5 < x'_6 < r'_3 < r'_4 < r'_5.$$

[3] See *Notation* at the beginning of the chapter.

Then, it (i.e. the move/assignment) is an f-assignment (see Figure 5.3) on $\{x'_0, x'_1 \dots x'_6\}$ with $f'(x) = (r'_1 - x)(r'_2 - x)(r'_3 - x)(r'_4 - x)(r'_5 - x)$ viz.

$$t' = \sum_{i=0}^{6} \frac{-f'(x'_i)}{\prod_{j\neq i}(x'_j - x'_i)} [\![x'_i]\!] \,.$$

Let $\Delta > 0$ denote a positive number. First, we use Lemma 83 and instead consider an f-assignment on $\{x_0, x_1 \dots x_6\}$ where $x_i = x'_i + \Delta$ with $f(x) = (r_1 - x)(r_2 - x) \dots (r_5 - x)$ where $r_i = r'_i + \Delta$ viz.

$$t = \sum_{i=0}^{6} \frac{-f(x_i)}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!] \,.$$

The lemma guarantees that the solution (as defined at the beginning of the chapter) to t and t' are the same. We decompose t into a sum of monomial assignments, i.e.

$$\begin{aligned} t = &\underbrace{\sum_{i=0}^{6} \frac{-r_1r_2r_3r_4r_5}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{I}} + \underbrace{\sum_{i=0}^{6} \frac{-\overbrace{(r_2r_3r_4r_5 + r_1r_3r_4r_5 + r_1r_2r_3r_5 + r_1r_2r_3r_4)}^{:=\alpha_1}(-x_i)}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{II}} \\ &+ \underbrace{\sum_{i=0}^{6} \frac{-\alpha_2(-x_i)^2}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{III}} + \underbrace{\sum_{i=0}^{6} \frac{-\alpha_3(-x_i)^3}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{IV}} \\ &+ \underbrace{\sum_{i=0}^{6} \frac{-\alpha_4(-x_i)^4}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{V}} + \underbrace{\sum_{i=0}^{6} \frac{-\alpha_5(-x_i)^5}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!]}_{\text{VI}} \end{aligned}$$

where α_l is the coefficient of $(-x)^l$ in $f(x)$. Since the total number of points in each assignment are 7, they are unbalanced (monomial assignments). Terms I, III and V each have an even powered monomial therefore they correspond to the aligned case. Their solutions, thus, can be read off directly from Proposition 89. Analogously, the remaining terms II, IV and VI each have an odd powered monomial therefore they correspond to the misaligned case. Their solutions, thus, can be read off directly from Proposition 90.

Outline—The Assembled Protocol

For completeness, we briefly outline how all the pieces fit together to give the full protocol, although in reverse. We describe the procedure in the language of TDPGs because, recall from Chapter 4 that, each step of the TDPG can be thought of as a short-hand to denote an exchange (and manipulation) of qubits between Alice and Bob granted the associated unitaries are known. As we have already done all the hard work in finding these unitaries[4], we can now proceed at this level of description. While describing the relation between TIPGs and TDPGs (i.e. Theorem 39) we deferred the proof to the appendix (i.e. Section A.3). As we need to go through this step (with minor modifications), to fully understand this procedure, the aforementioned is a prerequisite. An overall idea, however, may still be gleaned, solely from the following.

To be concrete, we use a bias $1/14$ game (see Figure 5.3; introduced earlier in Figure 2.10).

1. The first frame. This simply corresponds to the function $\frac{1}{2}([\![0, 1]\!] + [\![1, 0]\!])$.

[4] In this chapter we found the unitaries for f-assignments and in the previous chapter we found those corresponding to splits and merges.

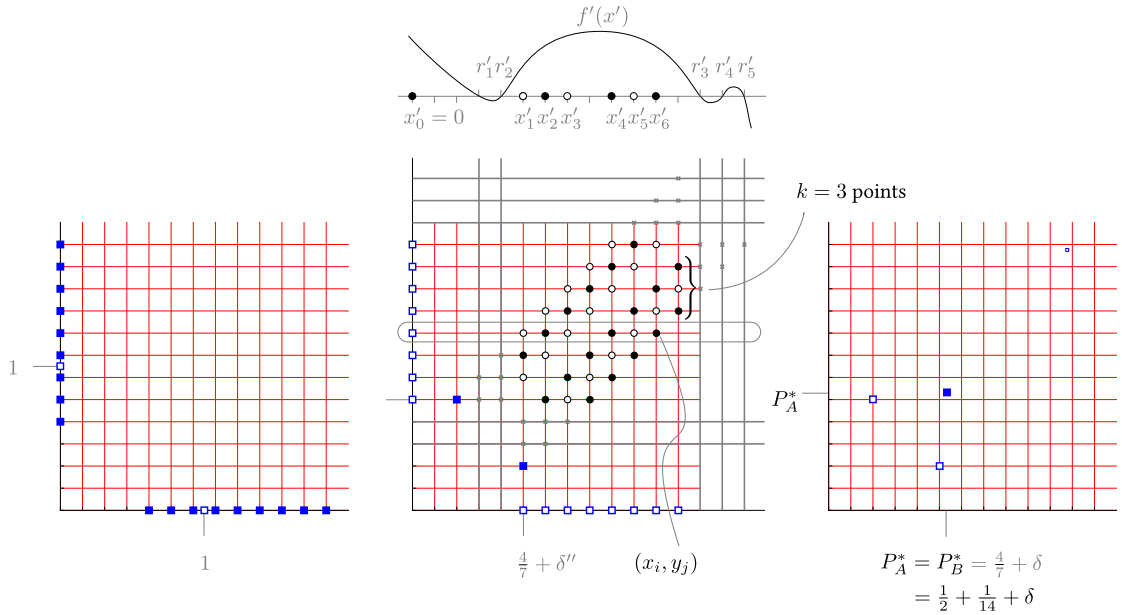


Figure 5.3: The TDPG (or equivalently, the reversed protocol) approaching bias $\epsilon(k = 3) = 1/14$ may be seen as proceeding in three stages, as illustrated by the three images (left to right). *First*, the initial points (indicated by unfilled squares) are split along the axes (indicated by the filled squares). *Second*, the points on the axes (unfilled squares) are transferred, via the ladder (indicated by the circles), into two final points (filled squares). *Third*, the two points from the previous step (unfilled squares) and the catalyst state (indicated, after being raised into one point, by the little unfilled box) are merged into the final point (filled box).
The *second* stage is illustrated by Mochon's TIPG (or more precisely, the so-called ladder) approaching bias $1/14$. Its typical move is highlighted. The weight of these points is given (up to a proportionality constant) by the f–assignment shown above. The roots of the polynomial correspond to the locations of the vertical lines and the location of the points in the graph is representative of the general construction.

2. The split. Deposit weights along the axis as specified by the TIPG of interest; more precisely, split the point $[\![0, 1]\!]$ into a set of points along the y–axis and analogously, split $[\![1, 0]\!]$ into a set of points along the x–axis, to match the distribution of points along the axis by the bias $1/14$ game.

3. The Catalyst State. Deposit a small amount of weight, δ_{catalyst}, at all the points that appear in the TIPG. This can be done, for instance, by raising (the x–coordinates) of the points which are along the y–axis, i.e. if the points along the axes are denoted as $\sum_i p_{\text{split},i} [\![0, y_i]\!]$ then raise them to obtain $\sum_i (p_{\text{split},i} - \delta_{\text{split},i}) [\![0, y_i]\!] + \sum_{i,j} \delta_{\text{catalyst}} [\![x_i, y_j]\!]$ where $\delta_{\text{catalyst}} > 0$ can be chosen to be arbitrarily small (and controls the number of repetitions in the next step) and the second sum is over points (x_i, y_j) which appear in the TIPG (excluding the axes[5]).

4. The Ladder.

   a) Denote the monomial decompositions of the valid functions by constituent valid functions. Globally scale these constituent valid functions sufficiently so that no negative weight appears when they are applied.

   b) Apply all the scaled down constituent horizontal valid functions.

   c) Apply all the scaled down constituent vertical valid functions.

[5] One needs to use the analogues procedure, i.e. use $\sum_i p_{\text{split},i} [\![x_i, 0]\!]$ as well for the one point of the TIPG which has a y–coordinate smaller than that of the points along the y–axis.

d) Repeat these two steps until essentially all the weight has been transferred from the axes into the two final points of the ladder[6].

Note that the unitaries corresponding to these constituent valid functions correspond to the solutions of the monomial assignments.

5. Raise and merge. Raise and merge the last two points into $(1-\delta')\left[\!\left[\frac{4}{7}+\delta'', \frac{4}{7}+\delta''\right]\!\right]$ where δ' represents the total weight used by the catalyst while δ'' comes from the truncation of the ladder, i.e. restricting the total number of points (in the ladder). Using the procedure introduced in the proof of Theorem 39 (see Section A.3), absorb the catalyst state (see Figure A.2) to obtain a single point $\left[\!\left[\frac{4}{7}+\delta, \frac{4}{7}+\delta\right]\!\right]$, i.e. $P_A^* = P_B^* = \frac{1}{2} + \frac{1}{14} + \delta$, where δ can be made arbitrarily small (but not zero) by making the catalyst state smaller and the ladder longer.

The actual protocol is just the reverse: it starts with a single point (corresponds to uncorrelated states) whose coordinates encode the cheating probabilities and ends with two points along the axis with equal weights (corresponds to the state $\frac{|AA\rangle+|BB\rangle}{\sqrt{2}}$).

[6]It would automatically become impossible to apply the moves once the weights on the axes becomes sufficiently small

III

Part

Secondary Contributions

6

Chapter

Elliptic Monotone Align (EMA) Algorithm

So far we have exclusively studied Mochon's point games. We now switch gears and construct a numerical algorithm that can generate the required unitary for any given Λ-valid function (see Definition 62). Note that corresponding to any WCF protocol with valid functions, one can find a WCF protocol with strictly valid functions (see Lemma 74). All strictly valid functions are Λ-valid for some finite Λ (see Lemma 71, Corollary 65). Thus we do not lose generality by restricting to Λ-valid functions.

§ 6.1 Canonical Forms Revisited

In this section we formalise the non-trivial constraint Equation (4.1) into two forms which we call the Canonical Projective Form (CPF) and the Canonical Orthogonal Form (COF). The CPF is always well defined but the corresponding COF may contain diverging eigenvalues. However since we restrict to Λ-valid functions, as we will see, the COF will also always be well defined. We need the COF as it is this that we use in the Elliptic Monotone Algorithm (EMA) algorithm.

6.1.1 The Canonical Projective Form (CPF) and the Canonical Orthogonal Form (COF)

We use the convention $p_{g_i}, p_{h_i} > 0$. This is important else in some of the statements one can find trivial counter-examples. Recall Theorem 75 which formally states the main result of Chapter 4. Note that the number of points initially, n_g, and finally, n_h, may differ. To facilitate further discussion we formalise the aforesaid condition into an object and its property. First, however, we define the following notation.

Definition 91 (Transition)**.** Consider two finitely supported functions $g, h : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$. A transition is defined as $g = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!] \to h = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$ where $[\![y]\!](x) := \delta_{xy}$ and $p_{g_i} > 0, p_{h_i} > 0$.

Definition 92 (Canonical Projective Form (CPF) for a give transition)**.** For a given transition the *Canonical Projective Form (CPF)* is given by the set of $m \times m$ matrices X_h, X_g, U, D and m dimensional vectors $|v\rangle$, $|w\rangle$ where

$$X_h := \sum_{i=1}^{n_h} x_{h_i} |h_i\rangle \langle h_i| , \; X_g := \sum_{i=1}^{n_g} x_{g_i} |g_i\rangle \langle g_i| ,$$

$$|w\rangle := \sum_{i=1}^{n_h} \sqrt{p_{h_i}} |h_i\rangle , \; |v\rangle := \sum_{i=1}^{n_g} \sqrt{p_{g_i}} |g_i\rangle ,$$

$$D := X_h - E_h U X_g U^\dagger E_h$$

and U is a unitary which satisfies

$$U |v\rangle = |w\rangle$$

for $E_h = \sum |h_i\rangle \langle h_i|$, orthonormal basis vectors $\left\{ |g_1\rangle, |g_2\rangle \dots \left|g_{n_g}\right\rangle, |h_1\rangle, |h_2\rangle \dots |h_{n_h}\rangle \right\}$, $m = n_g + n_h$.

Definition 93 (legal CPF)**.** A CPF is *legal* if $D \geq 0$.

In this language then our objective is to find a legal CPF for a given transition.

It *suffices to restrict to real unitaries viz. orthogonal matrices.* This will be justified in the next section but we already make this restriction in everything that follows (unless stated otherwise). In this section we try to reach an equivalence between a legal CPF and what we call the legal Canonical Orthogonal Form (COF).

The latter will be, roughly speaking, an inequality of the form $X_h - OX_gO^T \geq 0$ where $X_h = \text{diag}(x_{h_1}, x_{h_2} \dots, x_{h_{n_h}}, \xi, \xi \dots)$ and $X_g = \text{diag}(x_{g_1}, x_{g_2} \dots, x_{g_{n_g}}, 0, 0 \dots)$ for a large ξ. It is easy to see that if we can find an O that satisfies the COF for a given transition then the same O would satisfy the TEF inequality. It is almost trivial to note that a Λ valid function admit matrices of the COF form but we will show this later. Proving the other way, i.e. every legal CPF entails the corresponding COF must also be legal, is more non-trivial. Doing this requires handling the infinities and the matrix sizes more carefully. We only sketch an argument for this as we do not use it in the algorithm.

Definition 94 ((n, ξ) Canonical Orthogonal Form (COF) for a transition, ξ COF for a transition)**.** For a given transition and two numbers $n \geq \max(n_h, n_g)$, $\xi \geq \max(x_{h_1}, x_{h_2} \dots x_{h_{n_h}})$ an (n, ξ) *Canonical Orthogonal Form (COF)* is given by the set of $n \times n$ matrices X_h, X_g, O, D and vectors $|v\rangle$, $|w\rangle$ where

$$
\begin{aligned}
X_h &:= \text{diag}(x_{h_1}, x_{h_2} \dots, x_{h_{n_h}}, \xi, \xi \dots), \\
X_g &:= \text{diag}(x_{g_1}, x_{g_2} \dots, x_{g_{n_g}}, 0, 0 \dots), \\
|v\rangle &:= \sum_{i=1}^{n_g} \sqrt{p_{g_i}} \, |i\rangle \,, \\
|w\rangle &:= \sum_{i=1}^{n_h} \sqrt{p_{h_i}} \, |i\rangle \,, \\
D &:= X_h - OX_gO^T
\end{aligned}
$$

and the matrix O is orthogonal which satisfies

$$O \, |v\rangle = |w\rangle \,.$$

A ξ *Canonical Orthogonal Form (COF)* is an (n, ξ) COF with $n = n_h + n_g - 1$.

Definition 95 (n legal COF, legal COF)**.** An (n, ξ) COF is an n *legal COF* if $D \geq 0$ in the limit $\xi \to \infty$. A *legal COF* is a ξ COF such that $D \geq 0$ in the limit $\xi \to \infty$.

Imagine you found a legal COF corresponding to some transition. One can then sandwich D between a positive matrix as EDE to get

$$
\left[\begin{array}{c|ccc} X_h & & & \\ \hline & 1 & & \\ & & \ddots & \\ & & & 1 \end{array}\right] - \underbrace{\left[\begin{array}{ccc|ccc} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ \hline & & & \xi^{-1/2} & & \\ & & & & \ddots & \\ & & & & & \xi^{-1/2} \end{array}\right]}_{:=E} U X_g U^\dagger \left[\begin{array}{ccc|ccc} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ \hline & & & \xi^{-1/2} & & \\ & & & & \ddots & \\ & & & & & \xi^{-1/2} \end{array}\right] .
$$

Note that $D \geq 0 \iff EDE \geq 0$ because E is diagonal (which means one can write $EDE = (E\sqrt{D})(\sqrt{D}E)$ which in turn is of the A^TA form). From the legality of the COF, $D \geq 0$ in the limit $\xi \to \infty$ and in this limit E becomes a projector. After some relabelling (and appropriately expanding the space to $m = n_g + n_h$ dimensions) the inequality reduces to a CPF. This observation readily extends to the n legal case where $n \leq n_g + n_h$. It turns out that one can, and we show this later, always express an n' legal COF as an n legal COF with $n \leq n_g + n_h$ (in fact we can prove that $n \leq n_g + n_h - 1$). We have established the following statement.

Proposition 96. *Consider a transition. If there exists an n legal COF corresponding to it then there exists a legal CPF for the said transition.*

How about the reverse? Given a legal CPF can we find the corresponding n legal COF? We are given

$$D = \begin{bmatrix} X_h & & \\ & 0 & \\ & & \ddots \\ & & & 0 \end{bmatrix} - \underbrace{\begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix}}_{=E_h} U \begin{bmatrix} 0 & & \\ & X_g & \end{bmatrix} U^\dagger \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & 1 & \\ & & & 0 \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \geq 0.$$

Replacing the appended diagonal zeros in the first matrix (one containing X_h) with 1s yields an equivalent inequality. Next note that we can conjugate by a permutation matrix to get

$$\begin{bmatrix} 0 & & \\ & X_g & \end{bmatrix} = \tilde{U} \begin{bmatrix} X_g & & \\ & 0 & \end{bmatrix} \tilde{U}.$$

Finally we write the diagonal zeros in E_h as $1/\xi^{1/2}$ and use the reverse of the trick above to recover an m legal COF where recall $m := n_g + n_h$. This sketches the proof of the following statement.

Proposition 97. *Consider a transition. If there exists legal CPF corresponding to it then there exists an m legal COF for the said transition (where recall $m := n_g + n_h$).*

6.1.2 From EBM to EBRM to COF

We briefly summarise, at the cost of being redundant, how Aharonov et al. prove that valid functions are equivalent to the Expressible-By-Matrices (EBM) functions (assuming the operator monotones are on/the spectrum of the matrices is in $[0, \Lambda]$). They do this by showing that the set of EBM functions forms a convex cone K. Then they take the dual of this cone to get K^*. *This dual happens to be the set of operator monotone functions.* Then they find the bi-dual K^{**} and define the objects in this to be valid functions. They then show that $K = K^{**}$ which is to say that valid functions are equivalent to EBM functions. Note that all of this is assuming the aforesaid $[0, \Lambda]$ condition.

This is an extremely useful result because checking if a function is EBM is hard. Checking if a function is valid is a piece of cake because mathematical wizards have neatly characterised the set of operator monotone functions.

One can do even better. Instead of EBM functions, consider Expressible-By-Real-Matrices (EBRM) functions where the matrices are additionally restricted to be real. Let this set be given by K'. It turns out that its dual K'^* is also the set of operator monotone functions [14] viz. $K'^* = K^*$. Aharonov et al's proof for $K = K^{**}$ can be applied to the real case as is to get $K' = K^{**}$ (granted we assume the same $[0, \Lambda]$ condition).

Since Mochon's point games (and even the ones built later) use valid functions, the aforesaid simplification justifies why it suffices to restrict to real matrices.

We use the definition of Prob (Definition 24), EBM line transition (Definition 25), EBM function (Definition 48, Definition 49), Operator Monotone functions (Definition 56, Definition 57) and their characterisation (Lemma 59, Lemma 60), Λ valid functions (Definition 62) and finally its equivalence with EBM functions (Corollary 65).

Equivalence of EBM and EBRM

First we define EBRM transitions and EBRM functions similar to their EBM analogues except with the further restriction that the matrices and vectors involved are real.

Definition 98 (EBRM transitions)**.** Let $g, h : [0, \infty) \to [0, \infty)$ be two functions with finite supports. The transition $g \to h$ is EBRM if there exist two real matrices $0 \leq G \leq H$ and a (not necessarily normalised) vector $|\psi\rangle$ such that $g = \text{prob}[G, \psi]$ and $h = \text{prob}[H, \psi]$.

Definition 99 (K', EBRM functions; K'_Λ, EBRM functions on $[0, \Lambda]$)**.** A function $a : [0, \infty) \to \mathbb{R}$ with finite support is an EBRM function if the transition $a^- \to a^+$ is EBRM, where $a^+ : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $a^- : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ denote, respectively, the positive and the negative part of a ($a = a^+ - a^-$). We denote by K' the set of EBRM functions.

For any finite $\Lambda \in (0, \infty)$, a function $a : [0, \Lambda) \to \mathbb{R}$ with finite support is an EBRM function with support on $[0, \Lambda]$ if the transition $a^- \to a^+$ is EBRM with its spectrum in $[0, \Lambda]$, where a^+ and a^- denote, respectively, the positive and the negative part of a (again, $a = a^+ - a^-$). We denote by K'_Λ the set of EBRM functions with support on $[0, \Lambda]$.

Definition 100 (Real operator monotone functions)**.** A function $f : (0, \infty) \to \mathbb{R}$ is a real operator monotone if for all real matrices $0 \leq A \leq B$ we have $f(A) \leq f(B)$.

A function $f : (0, \Lambda) \to \mathbb{R}$ is a real operator monotone on $[0, \Lambda]$ if for all real matrices $0 \leq A \leq B$ with spectrum in $[0, \Lambda]$ we have $f(A) \leq f(B)$.

Lemma. *K'^*_Λ is the set of real operator monotones on $[0, \Lambda]$.*

Proof sketch. Aharonov et al. showed that K^*_Λ (which is, recall, the dual of the set of EBM functions on $[0, \Lambda]$) is the set of operator monotone functions on $[0, \Lambda]$ (see Lemma 3.9 of [3]). Their proof can be adapted here by restricting to real matrices which entails that K'^*_Λ is the set of real operator monotone functions on $[0, \Lambda]$. □

Lemma 101. *$K^*_\Lambda = K'^*_\Lambda$ and $K^* = K'^*$, i.e. the set of operator monotones on $[0, \Lambda]$ equals the set of real operator monotones on $[0, \Lambda]$ and the set of operator monotones equals the set of real operator monotones.*

Corollary 102. *$K'_\Lambda = K'^{**}_\Lambda = K^{**}_\Lambda = K_\Lambda$, i.e. the set of EBRM functions on $[0, \Lambda]$ = the set of Λ valid functions (dual of EBRM functions) = the set of Λ valid functions (dual of EBM functions) = the set of EBM functions on $[0, \Lambda]$.*

Corollary 103. *Any strictly valid function is EBRM.*

We now sketch the proof of Lemma 101. It is clear that the set of real operator monotones must contain the set of operator monotones because operator monotones are by definition required to work in the restricted real case as well. The set of real operator monotones might be bigger but that does not

happen to be the case. This is because we can encode an $n \times n$ hermitian matrix into a $2n \times 2n$ real symmetric matrix. This is achieved by replacing each complex number $\alpha + i\beta$ with the matrix

$$\alpha \begin{bmatrix} 1 & \\ & 1 \end{bmatrix} + \beta \begin{bmatrix} & -1 \\ 1 & \end{bmatrix}.$$

Note that the matrices have the exact same properties as 1 and i respectively. This corresponds to (after some permutation) writing a complex matrix $W = W_{\Re} + iW_{\Im}$ as a real symmetric

$$W' = \begin{bmatrix} W_{\Re} & -W_{\Im} \\ W_{\Im} & W_{\Re} \end{bmatrix}$$

where $W_{\Re}$ and $W_{\Im}$ are real. For a Hermitian $W^\dagger = W$ we must have $W_{\Re} = W_{\Re}^T$ and $W_{\Im} = -W_{\Im}^T$ which makes $W' = W'^T$ a symmetric matrix. The spectra of W and W' are the same. This is established most easily by looking at the diagonal decomposition. $W = USU^\dagger$ which would get transformed to $W' = U'S'U'^\dagger$. Since S is real S' is also real with doubly degenerate eigenvalues (except for the degeneracy already present in S). Thus if we have $0 \leq A \leq B$ then we would also have $0 \leq A' \leq B'$ as $A - B$ and $A' - B'$ would have the same spectrum where we used A' and B' to represent real symmetric analogues of the hermitian matrices A and B. The other way is trivial. Hence we have an equivalence which means that requiring a function to be operator monotone on complex matrices is the same as requiring it to be operator monotone on real symmetric matrices of twice the size (at most). This means that the set of real operator monotones is the same as the set of operator monotones.

EBRM to COF | Mochon's Variant

We just saw how we can reduce our problem from the set of EBM transitions to the set of EBRM transitions. We now show that each EBRM transition can be written in a standard form, which we call the Canonical Orthogonal Form (COF). The following is actually due to Mochon/Kitaev [28] but there was a minor mistake in one of the arguments which we have corrected. The interesting part is showing that one can always restrict the matrices to a certain size which in turn depends on the number of points involved in the transition.

Lemma 104. *For every EBRM transition $g \to h$ with spectrum in $[a, b]$ there exists an orthogonal matrix O, diagonal matrices X_h, X_g (with no multiplicities except possibly those of a and b) of size at most $n_g + n_h - 1$ such that*

$$O \underbrace{\begin{bmatrix} x_{g_1} & & & & \\ & \ddots & & & \\ & & x_{g_{n_g}} & & \\ & & & a & \\ & & & & \ddots \end{bmatrix}}_{:=X_g} O^T \leq \begin{bmatrix} x_{h_1} & & & & \\ & \ddots & & & \\ & & x_{h_{n_h}} & & \\ & & & b & \\ & & & & \ddots \end{bmatrix} = X_h,$$

and the vector $|\psi\rangle := \sum_{i=1}^{n_h} \sqrt{p_{h_i}} |i\rangle = \sum_{i=1}^{n_g} \sqrt{p_{g_i}} O |i\rangle$.

Proof. An EBRM entails that we are given $G \leq H$ with their spectrum in $[a, b]$ and a $|\psi\rangle$ such that

$$g = \text{Prob}[G, |\psi\rangle] = \sum_{i=1}^{n_g} p_{g_i}[x_{g_i}]$$

and

$$h = \text{Prob}[H, |\psi\rangle] = \sum_{i=1}^{n_h} p_{h_i}[x_{h_i}]$$

with $p_{g_i}, p_{h_i} > 0$ and $x_{g_i} \neq x_{g_j}$, $x_{h_i} \neq x_{h_j}$ for $i \neq j$ but the dimension and multiplicities can be arbitrary. First we show that one can always choose the eigenvectors $|g_i\rangle$ of G with eigenvalue x_{g_i} such that

$$|\psi\rangle = \sum_{i=1}^{n_g} \sqrt{p_{g_i}} |g_i\rangle .$$

Consider P_{g_i} to be the projector on the eigenspace with eigenvalue x_{g_i}. Note that

$$|g_i\rangle := \frac{P_{g_i} |\psi\rangle}{\sqrt{\langle\psi| P_{g_i} |\psi\rangle}}$$

fits the bill. Similarly we choose/define $|h_i\rangle$ so that

$$|\psi\rangle = \sum_{i=1}^{n_h} \sqrt{p_{h_i}} |h_i\rangle .$$

Consider now the projector onto the $\{|g_i\rangle\}$ space

$$\Pi_g = \sum_{i=1}^{n_g} |g_i\rangle \langle g_i| .$$

Note that this will not have all eigenvectors with eigenvalues $\in \{x_{g_i}\}$. Similarly we define

$$\Pi_h = \sum_{i=1}^{n_h} |h_i\rangle \langle h_i| .$$

We further define $G' := \Pi_g G \Pi_g + a(\mathbb{I} - \Pi_g)$ and $H' := \Pi_h H \Pi_h + b(\mathbb{I} - \Pi_h)$. These definitions are useful as we can show

$$G' \leq H'.$$

From $G = \Pi_g G \Pi_g + (\mathbb{I} - \Pi_g) G (\mathbb{I} - \Pi_g)$ we can conclude that $\Pi_g G \Pi_g + a(\mathbb{I} - \Pi_g) \leq G$. This entails $G' \leq G$. Using a similar argument one can also establish that $H \leq H'$. Combining these we get $G' \leq H'$.
Consider the projector

$$\Pi := \text{projector on span}\{\{|g_i\rangle\}_{i=1}^{n_g}, \{|h_i\rangle\}_{i=1}^{n_h}\}$$

and note that this has at most $n_g + n_h - 1$ dimension because $|\psi\rangle$ lives in the span of $\{|g_i\rangle\}$ and in the span of $\{|h_i\rangle\}$ so one of the basis vectors at least is not independent. Now note that

$$G'' := \Pi G' \Pi \leq \Pi H' \Pi =: H''$$

because we can always conjugate an inequality by a positive semi-definite matrix on both sides. Note also that $\Pi |\psi\rangle = |\psi\rangle$ which means the matrices and the vectors have the claimed dimension. We now establish that $\text{Prob}[H'', |\psi\rangle] = h$ and $\text{Prob}[G'', |\psi\rangle] = g$. For this we first write the projector tailored to the g basis as $\Pi = \Pi_g + \Pi_{g_\perp}$ where $\Pi_{g_\perp}$ is meant to enlarge the space to the span$\{h_i\}_{i=1}^{n_h}$. With this we evaluate

$$\begin{aligned} G'' &= (\Pi_g + \Pi_{g_\perp}) [\Pi_g G \Pi_g + a(\mathbb{I} - \Pi_g)] (\Pi_g + \Pi_{g_\perp}) \\ &= \Pi_g G \Pi_g + a \Pi_{g_\perp} . \end{aligned}$$

Manifestly then $\text{Prob}[G'', |\psi\rangle] = g$. By a similar argument one can establish the h claim. Note that that G'' and H'' have no multiplicities except possibly in a and b respectively. Thus we conclude we can always restrict to the claimed dimension and form. □

Corollary 105. *For every EBRM transition the corresponding COF is legal.*

The COF is of interest because we can use it to interpret our inequalities geometrically and use the tools thereof. We study this connection next.

§ 6.2 Ellipsoids

6.2.1 The inequality as containment of ellipsoids

We try to show that the matrix inequality of the form $0 \leq G \leq H$ can be geometrically viewed as the containment of a smaller ellipsoid inside a larger one.

Consider an unnormalised vector $|u\rangle = \sum_j u_j |h_j\rangle$ with $u_j \in \mathbb{R}$. Note that the set

$$\{|u\rangle \mid \langle u| X_h |u\rangle = 1\}$$

describes the surface of an ellipsoid where $X_h = \text{diag}(x_{h_1}, x_{h_2} \dots)$. This is easy to see as the constraint corresponds to

$$x_{h_1} u_1^2 + x_{h_2} u_2^2 + \dots = 1$$

which is of the form

$$\frac{u_1^2}{a_1^2} + \frac{u_2^2}{a_2^2} + \dots = 1$$

which, in turn, is the equation of an ellipsoid in the variables $\{u_i\}$ with axes $a_1 = 1/\sqrt{x_{h_1}}, a_2 = 1/\sqrt{x_{h_2}} \dots$. An inequality would correspond to points inside or outside the ellipsoid. It is also useful to note that if we start with some arbitrary (even unnormalised) vector $|u\rangle$ then the point on the ellipse along this direction are given by

$$\mathcal{E}_h(|u\rangle) = \frac{|u\rangle}{\sqrt{\langle u| X_h |u\rangle}}.$$

Finally, note that the set $\left\{|u\rangle \mid \langle u| U X_g U^\dagger |u\rangle = 1\right\}$ also corresponds to the equation of an ellipsoid with axes $\left\{1/\sqrt{x_{g_i}}\right\}$ except that it is rotated. This follows from the fact that if we use $|u'\rangle = U |u\rangle$ then the equation reduces to the standard form in the u'_i variables which can then be used to obtain u_is by the aforesaid relations which is a rotation. We can define a similar map from a vector $|u\rangle$ to a point on the rotated ellipse as

$$\mathcal{E}_g(|u\rangle) = \frac{|u\rangle}{\sqrt{\langle u| U X_g U^\dagger |u\rangle}}.$$

With this understanding in place we are ready to get a visual interpretation of our equation. The statement that

$$\begin{aligned}
& X_h - U X_g U^\dagger \geq 0 \\
\iff & \langle u| X_h |u\rangle - \langle u| U X_g U^\dagger |u\rangle \geq 0 && \forall\, |u\rangle \\
\iff & \langle u| U X_g U^\dagger |u\rangle \leq 1 && \forall\, \{|u\rangle \mid \langle u| X_h |u\rangle = 1\}
\end{aligned}$$

which in turn corresponds to the statement that every point denoted by $|u\rangle$ that is on the h ellipse must be on or inside the g ellipse. Note that if $\langle x_h\rangle - \langle x_g\rangle = 0$ then for $|u\rangle = |w\rangle$ the inequality saturates. This in turn means that even for $\mathcal{E}_h(|w\rangle)$ the inequality is saturated as it is the same vector

up to a scaling. The difference is that $\mathcal{E}_h(|w\rangle)$ represents a point on the h ellipsoid. Since the inequality is saturated it means that the ellipsoids must touch at this point. Thus $\mathcal{E}_g(|w\rangle) = \mathcal{E}_h(|w\rangle)$ which one can check explicitly as well.

The discussion so far can only give some intuition about the visualisation of our constraint equation. This intuition, as was explained in Subsection 1.3.3, can be efficiently used but it requires us to precisely specify the notion of curvature.

6.2.2 Convex Geometry Tools | Weingarten Map and the Support Function

Consider a curve in the plane. One can easily guess that the curvature must be related to the rate of change of tangents. This means we must use the second derivative. This can be generalised to arbitrary dimensions and in this case we obtain a matrix of the form $\partial_i \partial_j f$ for some function f which describes the curve. The eigenvalues of this matrix would tell us the curvature along the principal directions of curvature, given by the corresponding eigenvectors. It is possible to follow this idea through for an ellipsoid but the result becomes rather cumbersome as one must choose a coordinate system with its origin at the point of interest, aligned along the normal and re-express all the quantities of interest.

A concise way of evaluating the same is based on a mathematically sophisticated method applicable to all convex bodies. We state the result for the convex body of interest, an ellipsoid. For a more formal discussion (and related results) see Section D.1 in the Appendix.

For a normalised direction vector $|u\rangle$ the support function corresponding to an ellipsoid X is given by

$$h(u) = \sqrt{\langle u| X^{-1} |u\rangle} = \sqrt{\sum x_i^{-1} u_i^2}. \tag{6.1}$$

The derivative of the support function yields the point on the ellipsoid where the tangent plane corresponding to the direction $|u\rangle$ touches the said ellipsoid. It is

$$\partial_i h(u) = \frac{x_i^{-1} u_i}{h(u)}.$$

The most remarkable of all these is the fact is that

$$\partial_j \partial_i h(u) = \frac{1}{h(u)} \left(-\frac{x_j^{-1} x_i^{-1} u_i u_j}{h^2(u)} + x_i^{-1} \delta_{ij} \right) \tag{6.2}$$

contains as eigenvalues the radii of curvature at the aforesaid point and as eigenvectors the directions of principle curvature. In our discussion, for clarity, the preceding $1/h$ factor is ignored as it cancels out in the equations of interest anyway. If instead of the normal one knows the point at which one would like to evaluate this object then one can use the gradient to first find this normal and then apply the aforesaid. The normal at a point of contact $|c\rangle = \sum c_i |i\rangle$ is $|u(c)\rangle = \mathcal{N}\left(\sum x_i c_i |i\rangle\right)$. The results discussed here were deduced as special cases of those discussed in Section 2.5 of the book on convex bodies by R. Schneider [33].

We have stated the basic results needed to proceed with the description of our algorithm.

§ 6.3 Elliptic Monotone Align (EMA) Algorithm

Solving the weak coin flipping (WCF) problem can be reduced to finding explicit matrices for a given EBM transition $g = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!] \to h = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$ where g and h have disjoint support or, equivalently, for a given EBM function $a = h - g = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$. Here we describe our elliptic monotone align (EMA) algorithm, which runs by converting the given problem into the same problem of one less dimension iteratively until it is solved.

6.3.1 Notation

This subsection might appear to be particularly dry as we almost exclusively only introduce definitions; but it is a necessary evil. We try to motivate the definitions as we move along but things would not make perfect sense until one reaches the description and analysis of the algorithm itself.

At step k of the iteration, the transition $g \to h$ and the associated function $a = h - g$ used below are given by $g^{(k)} \to h^{(k)}$ and $a^{(k)}$ respectively. It remains fixed for the said step. We therefore do not write an explicit dependence on it in the following definitions. This is to facilitate the discussion of the iterative algorithm. We consider the extended real line $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty, -\infty\}$ with $1/\infty = -1/\infty := 0$. We also need the extended half line $\bar{\mathbb{R}}_{\geq} := \mathbb{R}_{\geq} \cup \{\infty\}$ and $\bar{\mathbb{R}}_{>} := \mathbb{R}_{>} \cup \{\infty\}$. These situations appear unavoidably in the analysis of certain transitions and correspond to one of the directions of the ellipsoids having infinite curvature. We use $\mathcal{N}(|\psi\rangle) := |\psi\rangle = \sqrt{\langle\psi|\psi\rangle}$. We usually denote by $[x_{\min}, x_{\max}]$ the smallest interval that contains $\operatorname{supp}(a)$. We call this interval the *support domain* for a. Similarly, we would refer to the smallest interval containing $\operatorname{supp}(g) \cup \operatorname{supp}(h)$ as the *transition support domain* for (the transition) $g \to h$. We use the variables $\chi, \xi \in \mathbb{R}$ to be such that they denote an interval $[\chi, \xi] \supseteq [x_{\min}, x_{\max}]$. As these χ and ξ would later be associated with an interval containing the spectrum of relevant matrices, we would refer to this interval as the *spectral domain*. The need for distinguishing the three is not hard to justify. A transition $g \to h$ can be such that both g and h have a term $p_k [\![x_k]\!]$ for some $p_k > 0$ and x_k. This term would be absent from $a = h - g$. Thus the transition support domain and the support domain would be different in general. One might object to this as we started with the assumption that g and h have disjoint support. The issue is that this assumption does not necessarily hold once the problem is reduced to a smaller instance of itself. As for the spectral domain, this is defined from hindsight, as we know we need to use COFs which (see Lemma 104) use matrices with spectra that would usually be larger than the transition support domain.

Recall from the discussion of Subsection 1.3.3 that we intend to use operator monotone functions to make the ellipsoids touch along a known direction. We already have a characterisation of operator monotone functions (see Lemma 60). The function $\lambda x/(\lambda + x)$ can be turned into $-1/(\lambda + x)$ by adding a constant (we will do this carefully shortly). Further, the characterisation we have expects the input matrices to have their spectrum in the range $[0, \Lambda]$. We must generalise this as this assumption can not be made for smaller instances of the same problem which appear in subsequent iterations. This motivates the following definitions.

Definition 106 (f_λ on (α, β))**.** $f_\lambda : (\alpha, \beta) \to \mathbb{R}$ is defined for $\lambda \in \mathbb{R} \backslash [-\beta, -\alpha]$ as

$$f_\lambda(x) := \frac{-1}{\lambda + x}.$$

Definition 107 (f_λ on $[\alpha, \beta]$)**.** $f_\lambda : [\alpha, \beta] \to \bar{\mathbb{R}}$ is defined for $\lambda \in \mathbb{R} \backslash (-\beta, -\alpha)$. For $\lambda \in \mathbb{R} \backslash [-\beta, -\alpha]$ we define

$$f_\lambda(x) := \frac{-1}{\lambda + x}.$$

For $\lambda = -\beta$ and $-\alpha$ we retain the same definition as above except when $x = \beta$ and α respectively in which case we define

$$f_{-\beta}(\beta) := \infty$$
$$f_{-\alpha}(\alpha) := -\infty.$$

Remark 108. Values for $f_{-\beta}(\beta)$ and $f_{-\alpha}(\alpha)$ are obtained by taking for x, respectively, the left limit (approaching from the left to β) and right limit (approaching from the right to α). Also note that the operator monotone $f(x) = x$ is not included in the aforesaid family of functions.

We had to define f_λ on the two intervals for technical reasons which we can't quite motivate here. We explicitly defined f_λ to be ∞ or $-\infty$ where it would be otherwise undefined (division by zero). These infinities, however, will only appear in the denominator in our algorithm.

Again, from the discussion of Subsection 1.3.3, we recall that we have to expand the smaller ellipsoid until it touches the larger ellipsoid. From Subsection 6.2.1 we can see that the ellipsoid corresponding X_h, a positive diagonal matrix, is smaller than the one corresponding to γX_h for $0 < \gamma < 1$. The X_h matrix would correspond to a function h. What would the corresponding function be for γX_h? The following definition of h_γ formalises the answer. We also introduce l_γ which helps us check the validity condition for a transition (similar to Definition 62 and Corollary 65). The basic idea is to take the inner product (sum over the finite support of a) of the function a with a given operator monotone. If this is positive for every operator monotone, then the function a is valid. From hindsight, since we already know the characterisation of these operator monotones, we define $l_\gamma(\lambda)$ to be this inner product which must be positive, labelling the operator monotone by λ and encoding the stretching of the h ellipsoid into γ. This plays a crucial role in our algorithm as we have to make sure we use the right stretching, γ, without actually knowing the ellipsoids completely. We do not expect the details of these statements to be clear just yet but we hope the following definitions appear reasonable.

Definition 109 ($l_\gamma, l_\gamma^1, a_\gamma$)**.** Consider the transition $g \to h$ and let $a = h - g$. For $\gamma \in (0, 1]$ we define the finitely supported functions $h_\gamma : \mathbb{R} \to \mathbb{R}_{\geq}$ and $a_\gamma(x) : \mathbb{R} \to \mathbb{R}$ as

$$h_\gamma(x) := h(x/\gamma)$$
$$a_\gamma(x) := h_\gamma(x) - g(x).$$

Let $S_\gamma = [x_{\min}(\gamma), x_{\max}(\gamma)]$ be the support domain of a_γ. We define $l_\gamma : \mathbb{R} \backslash [-x_{\max}(\gamma), -x_{\min}(\gamma)] \to \mathbb{R}$

$$l_\gamma(\lambda) := \sum_{x \in \text{supp}(a_\gamma)} a_\gamma(x) f_\lambda(x)$$

where f_λ is defined on S_γ.

We define

$$l_\gamma^1 := \sum_{x \in \text{supp}(a_\gamma)} a_\gamma(x) x.$$

Remark 110. h_γ and g might have overlapping support for certain values of γ which justifies the terminology distinguishing support domain and spectral support domain (introduced at the beginning of the section).

We now define a sort of indicator function, m, which tells us, given the transition $g \to h$, if the transition corresponding to the scaled ellipsoid $g \to h_\gamma$ is valid. There are some extra parameters this function needs. Consider the spectrum of the matrices which make this transition EBRM (they must be EBRM if they are valid, similar to Corollary 65). These parameters encode the interval in which this spectrum must be contained.

Definition 111 ($m(\gamma, \chi, \xi)$)**.** We define $m : ((0,1], \mathbb{R}, \mathbb{R}) \to \{0,1\}$ to be

$$m(\gamma, \chi, \xi) := \begin{cases} 0 & \text{if any of the following root conditions hold} \\ 1 & \text{else.} \end{cases}$$

where the first root condition is satisfied if there exists a $\lambda \in \mathbb{R}\backslash(-\xi, -\chi)$ such that $l_\gamma(\lambda) = 0$, and the second root condition is satisfied if $l_\gamma^1 = 0$.

As we are dealing with different representations of the same object, we define a relation between the matrix instance of the problem (which involves matrices) and the function instance thereof (which involves transitions and functions). The matrix instance contains all the information needed and so in the discussion of the algorithm we pack everything into a matrix instance to keep things palpable. The reader can glance through the following and later refer to them when they are used.

Definition 112 (Matrix Instance, $\underline{\mathrm{X}} \to$ Function Instance, $\underline{\mathrm{x}}$)**.** For a *Matrix Instance* defined to be the tuple $\underline{\mathrm{X}} := (X_h, X_g, |w\rangle, |v\rangle)$ where X_h, X_g are diagonal matrices and $|w\rangle, |v\rangle$ are vectors on $\mathbb{R}^n$ for some n with equal norm, i.e. $\langle w|w\rangle = \langle v|v\rangle$, we define the *Function Instance* to be the tuple $\underline{\mathrm{x}} : (g, h, a)$ where $h = \mathrm{Prob}[X_h, |w\rangle]$, $g = \mathrm{Prob}[X_g, |v\rangle]$ and $a = h - g$.

Definition 113 (Attributes of the Function Instance, $\underline{\mathrm{x}}$)**.** For a given tuple $\underline{\mathrm{x}} := (g, h, a)$ as defined in Definition 112 we define the attributes $n_h, n_g, \{p_{g_i}\}, \{p_{h_i}\}, \{x_{g_i}\}, \{x_{h_i}\}$ as they appear by declaring $g \to h$ to be a transition, i.e.,

- n_h as the number of times h is non-zero,
- n_g as the number of times g is non-zero,
- $\{p_{h_i}\}, \{x_{h_i}\}, \{p_{g_i}\}, \{x_{g_i}\}$ implicitly as

$$h = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!], \; g = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$$

  (for $p_{h_i}, p_{g_i} > 0$).

The *support domain* for a is denoted by $[x_{\min}, x_{\max}]$, i.e., the attributes $x_{\min}, x_{\max}$ are defined to be such that $[x_{\min}, x_{\max}]$ is the smallest interval containing $\mathrm{supp}(a)$.

Remark 114. Note that $x_{\min}$ and $x_{\max}$ may not be $x_{\min} = \min\{\{x_{h_i}\}, \{x_{g_i}\}\}$ and $x_{\max} = \max\{\{x_{h_i}\}, \{x_{g_i}\}\}$ respectively because there can be cancellations in the evaluation of $h - g = a$.

Definition 115 (Attributes of the Matrix Instance, $\underline{\mathrm{X}}$)**.** We associate the following with a matrix instance.

- *Spectral domain*: For a tuple $\underline{\mathrm{X}}$ as defined in Definition 112 we denote the *spectral domain* by $[\chi, \xi]$ where the attributes χ, ξ are such that $[\chi, \xi]$ is the smallest interval containing $\mathrm{spec}\{X_g \oplus X_h\}$.
- *Solution*: We say that O is a *solution* to the matrix instance $\underline{\mathrm{X}} = (X_h, X_g, |w\rangle, |v\rangle)$ if $X_h \geq O X_g O^T$ and $O|v\rangle = |w\rangle$.
- *Notation*: With respect to a standard orthonormal basis $\{|i\rangle\}$, we use the *notation* $X_h := \sum_{i=1}^k y_{h_i} |i\rangle\langle i|$, $X_g := \sum_{i=1}^k y_{g_i} |i\rangle\langle i|$, $|w\rangle := \sum_{i=1}^k \sqrt{q_{h_i}} |i\rangle$, $|v\rangle := \sum_{i=1}^k \sqrt{q_{g_i}} |i\rangle$.

Remark 116. We index the Matrix Instance and the corresponding Function Instance as

$$\underline{\mathrm{X}}^{(k)} = \left(X_h^{(k)}, X_g^{(k)}, \left|w^{(k)}\right\rangle, \left|v^{(k)}\right\rangle\right)$$

and

$$\underline{\mathrm{X}}^{(k)} \to \underline{\mathrm{x}}^{(k)} = \left(h^{(k)}, g^{(k)}, a^{(k)}\right)$$

respectively. The associated attributes are implicitly assumed to be correspondingly indexed, e.g., as $\chi^{(k)}, \xi^{(k)}$ and $n_h^{(k)}, n_g^{(k)}, x_{\min}^{(k)}, x_{\max}^{(k)}$.

Remark 117. We introduce two different symbol sets p, x and q, y as it allows us to describe the proof more neatly by allowing two ways of indexing the same object. We use p, x for $\underline{\mathrm{x}}$ and q, y for $\underline{\mathrm{X}}$ which are essentially the same.

6.3.2 Lemmas for EMA

With the notation in place, we can now state and prove some results which we would need in our algorithm. We do this in three steps. First, we generalise the results obtained by Aharonov et al. about operator monotones and their relation with EBM functions. This is the workhorse of our algorithm. Second, we prove some results which formalise our intuitive notion of tightening—stretching the smaller h ellipsoid until it touches the larger g ellipsoid. Finally, we prove a generalisation thereof in the case where the curvature of the smaller h ellipsoid becomes infinite.

For a first reading, it might be better to focus on the statements, and come back to the proofs after reading the algorithm.

6.3.2.1 Generalisations

Keep the bigger picture, Figure 6.1, in mind to retain a sense of direction. Our main objective here would be twofold. First, we wish to generalise Corollary 102 from being restricted to matrices with their spectrum in $[0, \Lambda]$ to being applicable for matrices with their spectrum in $[\chi, \xi]$. Second, we wish to extend the result from valid functions to valid transitions, including the case of overlapping support.

To establish the first, our strategy would be to find a relation between $[0, \Lambda]$ valid functions and $[\chi, \xi]$ valid functions (which we will define carefully soon) and then a relation between $[0, \Lambda]$ EBRM functions and $[\chi, \xi]$ EBRM functions. Then we use the link between $[0, \Lambda]$ valid and $[0, \Lambda]$ EBRM functions to establish the equivalence of $[\chi, \xi]$ valid functions and $[\chi, \xi]$ EBRM functions. Along the way we sharpen our understanding of operator monotone functions which should make the definitions of f_λ, l and m (see Definition 109 and Definition 111) obvious. The second objective can be met with by a single, albeit, slightly long argument.

Let us start with extending our definition of the Canonical Orthogonal Form to accommodate matrices with their spectrum in $[\chi, \xi]$.

Definition 118 (Canonical Orthogonal Form (COF) with spectrum in $[\chi, \xi]$)**.** For a given transition $g \to h$ let $[\chi, \xi]$ be such that it contains $\mathrm{supp}(g)$ and $\mathrm{supp}(h)$. We define the Canonical Orthogonal Form (COF) with its spectrum in $[\chi, \xi]$ by the set of $n \times n$ matrices X_h, X_g, O, D and vectors $|v\rangle, |w\rangle$ where

$$X_h := \mathrm{diag}\{x_{h_1}, x_{h_2} \ldots, x_{h_{n_h}}, \xi, \xi \ldots\},$$

$$X_g := \mathrm{diag}\{x_{g_1}, x_{g_2} \ldots, x_{g_{n_g}}, \chi, \chi \ldots\},$$

$$|v\rangle := \sum_{i=1}^{n_g} \sqrt{p_{g_i}}\, |i\rangle\,,$$

$$|w\rangle := \sum_{i=1}^{n_h} \sqrt{p_{h_i}} \, |i\rangle \, ,$$

$$D := X_h - O X_g O^\dagger,$$

the matrix O is orthogonal which satisfies

$$|v\rangle = O \, |w\rangle$$

and $n = n_g + n_h - 1$.

Definition 119 (Legal COF with spectrum in $[\chi, \xi]$)**.** A COF with spectrum in $[\chi, \xi]$ is legal if $D \geq 0$.

We obviously need to generalise the notion of operator monotone functions to the range $[\chi, \xi]$ as well to achieve our first objective.

Definition 120 (Operator monotone functions on $[\chi, \xi]$)**.** A function $f : [\chi, \xi] \to \mathbb{R}$ is operator monotone on $[\chi, \xi]$ if for all real symmetric matrices H, G with $\text{spec}(H \oplus G) \in [\chi, \xi]$ and $H \geq G$ we have $f(H) \geq f(G)$.

What happens if we try to shift/translate the interval on which an operator monotone is defined? This is a natural question to ask, an answer to which would also directly relate our new definition to the previous one.

Claim 121. $f(x)$ is an operator monotone function on $[\chi, \xi]$ if and only if $f'(x') = f(x' - x_0)$ is an operator monotone function on $[\chi + x_0, \xi + x_0]$.

Proof. Consider real symmetric matrices $H \geq G$ with $\text{spec}(H \oplus G) \in [\chi, \xi]$ and let $f(x)$ be operator monotone on $[\chi, \xi]$. We must consider $f'(x') = f(x = x' - x_0)$ which is the same as $f'(x + x_0) = f(x)$. We show that f' is an operator monotone on $[\chi + x_0, \xi + x_0]$. Note that $H' := H + x_0 \mathbb{I}$ and $G' := G + x_0 \mathbb{I}$ are such that $H' \geq G'$ and $\text{spec}(H' \oplus G') \in [\chi + x_0, \xi + x_0]$. Note that $f'(H') = f(H)$ and $f'(G') = f(G)$ because

$$\begin{aligned} f'(H') &= f'(H + x_0 \mathbb{I}) \\ &= O_h f'(H_d + x_0 \mathbb{I}) O_h^T \\ &= O_h f(H_d) O_h^T \\ &= f(H) \end{aligned}$$

and similarly for G where $H = O_h H_d O_h^T$ for O_h orthogonal and H_d diagonal. Since f is operator monotone on $[\chi, \xi]$ we have $f(H) \geq f(G)$ which entails $f'(H') \geq f'(G')$. Since this holds for all H', G' with their $\text{spec}(H' \oplus G') \in [\chi + x_0, \xi + x_0]$ we can conclude that f' is an operator monotone on $[\chi + x_0, \xi + x_0]$. The other way follows by setting $\chi + x_0$ to χ, $\xi + x_0$ to ξ, x_0 to $-x_0$ but since all these were arbitrary to start with, the reasoning goes through unchanged. □

We now note that from the characterisation of operator monotone functions we initially had (see Lemma 60), we can construct one which is easier to shift/translate (in the aforesaid sense).

Corollary 122 (Characterisation of operator monotone functions on $[0, \Lambda]$)**.** *Any operator monotone function $f : [0, \Lambda] \to \mathbb{R}$ can be written as*

$$f(x) = c_0 + c_1 x - \int \frac{1}{\lambda + x} d\tilde{\omega}(\lambda)$$

with the integral ranging over $\lambda \in (-\infty, -\Lambda) \cup (0, \infty)$ satisfying $\int \frac{1}{\lambda(1+\lambda)} d\tilde{\omega}(\lambda) < \infty$.

Proof. Consider the characterisation given in Lemma 60 according to which we had $f(x) = c_0' + c_1 x + \int \frac{\lambda x}{\lambda + x} d\omega(\lambda)$ with $\int \frac{\lambda}{1+\lambda} d\omega(\lambda) < \infty$. We can write

$$\begin{aligned} f(x) &= c_0' + c_1 x + \int \left(\lambda - \frac{\lambda^2}{\lambda + x} \right) d\omega(\lambda) \\ &= c_0 + c_1 x - \int \frac{\lambda^2 d\omega(\lambda)}{\lambda + x} \end{aligned}$$

where with $d\tilde{\omega} = \lambda^2 d\omega(\lambda)$ we obtain the claimed form. Note that the finiteness of $\int \frac{\lambda}{1+\lambda} d\omega$ is necessary to conclude that $c_0 = c_0' + \int \frac{\lambda}{1+\lambda} d\omega$ is also finite. □

Observe that this form also makes it easier for us to handle any divergences as there is only the denominator one has to deal with.

This can now be shifted/translated to allow for a characterisation of our shifted/translated operator monotones.

Corollary 123 (Characterisation of operator monotone functions on $[\chi, \xi]$)**.** *Any operator monotone function $f' : [\chi, \xi] \to \mathbb{R}$ can be written as*

$$f'(x') = c_0' + c_1' x' - \int \frac{1}{\lambda' + x'} d\tilde{\omega}'(\lambda')$$

with the integral ranging over $\lambda' \in (-\infty, -\xi) \cup (-\chi, \infty)$ satisfying $\int \frac{1}{(\lambda'+\chi)(1+\lambda'+\chi)} d\tilde{\omega}'(\lambda') < \infty$.

Proof. We follow the convention that $x' \in [\chi, \xi]$ while the unprimed $x \in [0, \xi - \chi]$. From Claim 121 we know that $f(x)$ is operator monotone on $[0, \xi - \chi]$ if and only if $f'(x') = f(x' - \chi)$ is operator monotone on $[\chi, \xi]$ where $x' = x + \chi$. Since we already have a characterisation for $f(x)$ we can characterise $f'(x')$ as $f(x' - \chi)$. From Corollary 122 we have

$$\begin{aligned} f'(x') &= c_0 + c_1(x' - \chi) - \int \frac{d\tilde{\omega}(\lambda)}{\lambda + x' - \chi} \\ &= c_0' + c_1 x' - \int \frac{d\tilde{\omega}'(\lambda')}{\lambda' + x'} \end{aligned}$$

where $\lambda' = \lambda - \chi$. Since we had $\lambda \in (-\infty, -(\xi - \chi)) \cup (0, \infty)$ it entails $\lambda' \in (-\infty, -\xi) \cup (-\chi, \infty)$. The condition on the integral $\int \frac{d\tilde{\omega}(\lambda)}{\lambda(\lambda+x)} < \infty$ can be expressed in terms of λ' as $\int \frac{d\tilde{\omega}'(\lambda')}{(\lambda'+\chi)(1+\lambda'+\chi)} < \infty$ with $d\tilde{\omega}'(\lambda') = d\tilde{\omega}(\lambda' + \chi)$. With $c_1 = c_1'$ and $c_0' = c_0 - c_1\chi$ we obtain the claimed form. □

We now generalise Definition 62 to $[\chi, \xi]$–valid functions[1].

Definition 124 ($[\chi, \xi]$–valid function)**.** A finitely supported function $a : \mathbb{R} \to \mathbb{R}$ with $\text{supp}(a) \in [\chi, \xi]$ is $[\chi, \xi]$–valid if for every operator monotone function f on $[\chi, \xi]$ we have $\sum_{x \in \text{supp}(a)} a(x) f(x) \geq 0$.

Remark 125. Since in Corollary 123 $d\tilde{\omega}'$ is a measure, to establish $[\chi, \xi]$ validity of functions, it would suffice to restrict our attention to operator monotones $f'(x') = x'$, $f'(x') = -\frac{1}{\lambda'+x'}$ with $x' \in [\chi, \xi]$, $\lambda' \in (-\infty, -\xi) \cup (-\chi, \infty)$.

By shifting/translating the characterisation of operator monotone functions we can shift/translate valid functions as well.

[1] If you spot a (χ, ξ)–valid function, take it to mean a $[\chi, \xi]$–valid function; it is an oversight.

Corollary 126 ($a(x)$ is $[\chi,\xi]$–valid $\iff$ $a(x'-x_0)$ is $[\chi+x_0,\xi+x_0]$–valid)**.** *A finitely supported function $a:\mathbb{R}\to\mathbb{R}$ with supp$(a)\in[\chi,\xi]$ is $[\chi,\xi]$–valid if and only if the function $a'(x'):=a(x'-x_0):\mathbb{R}_{\geq}\to\mathbb{R}$ is $[\chi-x_0,\xi-x_0]$–valid.*

Proof. *a is $[\chi,\xi]$ valid entails $\sum_{x\in supp(a)}a(x)f(x)\geq 0$ for all f operator monotone on $[\chi,\xi]$. We can write the sum as $\sum a(x'-x_0)f(x'-x_0)\geq 0$. Using Claim 121 we note that $f'(x')=f(x'-x_0)$ is operator monotone on $[\chi+x_0,\xi+x_0]$. For $a'(x')=a(x'-x_0)$ we thus have $\sum a'(x')f'(x')\geq 0$ which means $a'(x')$ is a $[\chi+x_0,\xi+x_0]$–valid function. The other way follows similarly.* □

In accordance with our strategy, we have established a relation between $[0,\Lambda]$–valid functions and (χ,ξ) valid functions (in fact we have a more general result). We now proceed with establishing its analogue for EBRM functions.

Definition 127 (EBRM on $[\chi,\xi]$)**.** A finitely supported function $a:\mathbb{R}\to\mathbb{R}$ is EBRM on $[\chi,\xi]$ if there exist real symmetric matrices $H\geq G$ with their spectrum in $[\chi,\xi]$ and a vector $|w\rangle$ such that $a=\text{Prob}[H,|w\rangle]-\text{Prob}[G,|w\rangle]$.

Corollary 128 ($a(x)$ is EBRM on $[\chi,\xi]$ $\iff$ $a(x+\chi)$ is EBRM on $[0,\xi-\chi]$)**.** *A finitely supported function $a:\mathbb{R}\to\mathbb{R}$ with supp$(a)\in[\chi,\xi]$ is EBRM on $[\chi,\xi]$ if and only if the function $a'(x):=(x+\chi):\mathbb{R}_{\geq}\to\mathbb{R}$ is EBRM on $[0,\xi-\chi]$.*

Proof. If a is EBRM on $[\chi,\xi]$ it follows that there exist real symmetric matrices with $H\geq G$ and a vector $|w\rangle$ such that spec$[H\oplus G]\in[\chi,\xi]$ and $a=\text{Prob}[H,|w\rangle]-\text{Prob}[G,|w\rangle]$. Clearly, $H':=H-\chi\mathbb{I}\geq G-\chi\mathbb{I}=:G'$ and $a'(x)=\text{Prob}[H',|w\rangle]-\text{Prob}[G',|w\rangle]=a(x+\chi)$ with spec$[H'\oplus G']\in[0,\xi-\chi]$. This means a' is EBRM on $[0,\xi-\chi]$. The other way follows similarly. □

We have done all the hard work for meeting the first objective. We now simply combine our results so far to prove the desired equivalence.

Lemma 129 ($a(x)$ is $[\chi,\xi]$–valid function $\iff$ $a(x)$ is EBRM on $[\chi,\xi]$)**.** *A finitely supported function $a:\mathbb{R}\to\mathbb{R}$ with supp$(a)\in[\chi,\xi]$ being $[\chi,\xi]$ valid is equivalent to it being EBRM on $[\chi,\xi]$.*

Proof. From Corollary 126 we know that $a(x)$ being $[\chi,\xi]$ valid is equivalent to $a(x+\chi)$ being $\Lambda=\xi-\chi$ valid. From Corollary 102 we know that $a(x+\chi)$ is equivalently EBRM on $[0,\xi-\chi]$. Finally using Corollary 128 we know that $a(x+\chi)$ being EBRM on $[0,\xi-\chi]$ is equivalent to $a(x)$ being EBRM on $[\chi,\xi]$. □

The second objective will be achieved in a single shot.

Lemma 130 (EBRM function $\iff$ EBRM transition even with common support)**.** *If we write an EBRM function a with spectrum in $[\chi',\xi']$ as $a=h-g$ with $h,g:\mathbb{R}_{\geq}\to\mathbb{R}_{\geq}$ which may have common support then $g\to h$ is an EBRM transition with spectrum in $[\chi,\xi]$ and with (the smallest) matrix size (at most) n_g+n_h-1 where $[\chi,\xi]$ is the smallest interval containing $[\chi',\xi']$ and supp$(h)\cup$ supp(g).*

Conversely, if $g\to h$ is an EBRM transition with spectrum in $[\chi,\xi]$ with $h,g:\mathbb{R}_{\geq}\to\mathbb{R}_{\geq}$ which may have common support then $a=h-g$ is an EBRM function with its spectrum in $[\chi,\xi]$ (the smallest) matrix size at most n_g+n_h-1.

Proof. To prove the first statement we write $a = a^+ - a^-$ where $a^+ = \sum_{i=1}^{n'_h} p'_{h_i} [\![x_{h_i}]\!], a^- = \sum_{i=1}^{n'_g} p'_{g_i} [\![x_{g_i}]\!]$, for $a^+, a^- : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$, represent the positive and the negative parts of a. Note that a^+ and a^- by virtue of this definition can't have any common support. Consider $\Delta = \sum_{i=1}^{n_\Delta} c_i [\![x_i]\!] : \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ to be such that $h = a^+ + \Delta$ and $g = a^- + \Delta$. This is always the case because $h - g = a$. Consider the case where $\text{supp}(\Delta) \cap \text{supp}(a) = \emptyset$. In this case $n_g = n'_g + n_\Delta$ and $n_h = n'_h + n_\Delta$. Since a is an EBRM function we have a legal COF, viz $O' X'_g O'^T \leq X'_h$ and $|w'\rangle = O' |v'\rangle$, of dimension $(n' = n'_g + n'_h - 1)$ from Lemma 104. To obtain the matrices corresponding to $g \to h$ we expand the space to $n = n_g + n_h - 1$ dimensions and define $X_g = X'_g \oplus X$, $X_h = X'_h \oplus X$, $O = O' \oplus \mathbb{I}$, $|v\rangle = |v'\rangle + \sum_{i=n'}^{n} \sqrt{c_{i+1-n'}} |i\rangle$ where $X = \text{diag}\{x_1, x_2 \ldots x_{n_\Delta}\}$. This is just an elaborate way of adding the points in Δ to the matrices and the vectors in such a way that the part corresponding to Δ remains unchanged. The other cases can be similarly demonstrated with the only difference being in the relation between n_g, n'_g and n_h, n'_h. Suppose Δ is non-zero only at one point. If Δ adds a point where a^- had a point then it does not contribute to increasing the number of points in g that is $n_g = n'_g$ but it does increase the number in h that is $n_h = n'_h + 1$. This means that we have one extra dimension to find the matrices certifying $g \to h$ is EBRM which is precisely what is needed to append that extra idle point as described above. Similarly one can reason for adding a point where a^+ had a point and finally extend it to the most general case of $\text{supp}(\Delta) \cap \text{supp}(a) \neq \emptyset$ which may involve multiple points.

We now prove the converse. Since $g \to h$ is an EBRM transition from Lemma 104 we know that it admits a legal COF, that is $O X_g O^T \leq X_h$ and $O |v\rangle = |w\rangle$ with dimension $n_g + n_h - 1$. To be able to show that $a = h - g = a^+ - a^-$ (where a^+ and a^- are again the positive and negative part of a) is an EBRM function it suffices to show that a is a valid function. This follows directly from the COF and operator monotones as $O f(X_g) O^T \leq f(X_h)$ implies $\langle v| f(X_g) |v\rangle \leq \langle w| f(X_h) |w\rangle$ which in turn is $\sum h(x) f(x) - \sum g(x) f(x) \geq 0$ and that is the same as $\sum a(x) f(x) \geq 0$ for all f operator monotone on the spectrum of $X_h \oplus X_g$, viz. a is valid. From Lemma 129 we conclude that a is also EBRM with size at most $n_g + n_h - 1$ (actually we can make a stronger statement by saying the size should be at most $n'_g + n'_h - 1$ where $|\text{supp}(a^+)| = n'_h$ and $|\text{supp}(a^-)| = n'_g$). □

This completes the first, and longest, step of our groundwork for discussing the algorithm. Our achievement so far has been schematized in Figure 6.1.

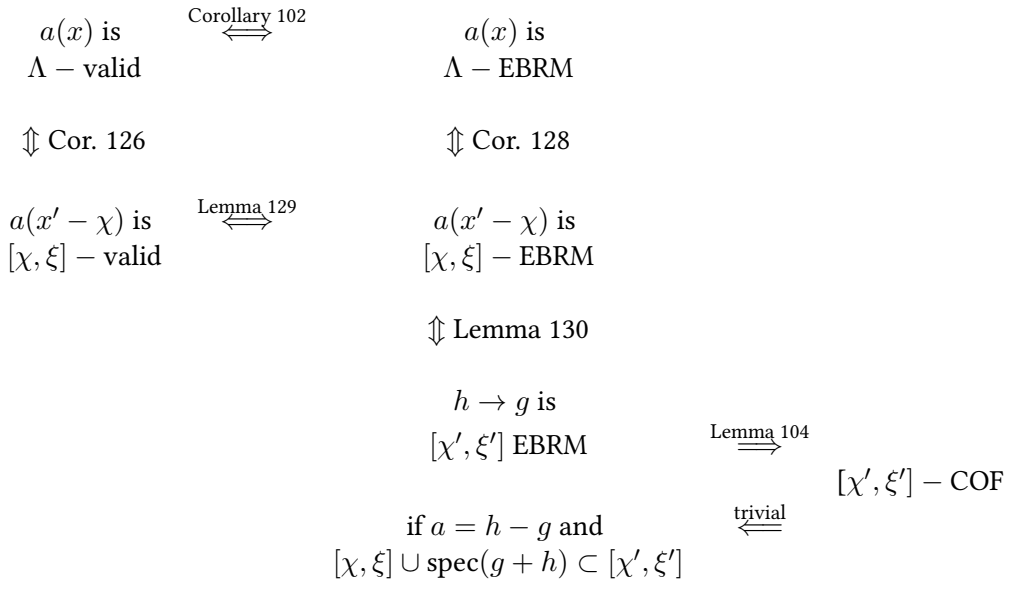


Figure 6.1: Generalisation schematized.

6.3.2.2 For the finite part

For the second step, we state the following fact (see [8]). The proof of this statement is interesting in its own right but here we only state it and use it for terminating our recursive algorithm.

Fact 131 (Weyl's Monotonicity Theorem). *If H is positive semi-definite and A is Hermitian then $\lambda_j^{\downarrow}(A+H) \geq \lambda_j^{\downarrow}(A)$ for all j where $\lambda_j^{\downarrow}(M)$ represents the j^{th} largest eigenvalue of the Hermitian matrix M.*

Corollary 132. *If $H \geq G$ then $\lambda_j^{\downarrow}(H) \geq \lambda_j^{\downarrow}(G)$ for all j.*

At some point, if the algorithm reaches a point where there is no vector constraint (that is the vector $|v\rangle = |w\rangle = 0$) then one can use the aforesaid result to conclude that the solution, orthogonal matrix O, must be a permutation matrix (this will become clear later, we mention it to motivate the relevance of the result).

We now state a continuity condition which we subsequently use to establish that when we stretch the h ellipsoid, there would always exist the perfect amount of stretching that makes the h ellipsoid just touch the g ellipsoid. The non-triviality here is that we have to conclude this without fully knowing the ellipsoids.

Claim 133 (Continuity of l). Let $[x_{\min}, x_{\max}]$ be the smallest interval containing $\text{supp}(a)$. $l(\lambda)$ is continuous in the intervals $\lambda \in (-x_{\min}, \infty]$ and $\lambda \in [-\infty, -x_{\max})$ (see Definition 109).

Proof. Since $l(\lambda)$ is just a rational function of λ it suffices to show that the denominator doesn't become zero in the said range. The roots of the denominator are of the form $\lambda + x = 0$ for $x \in \{\{x_{g_i}\}, \{x_{h_i}\}\}$. Hence the largest root will be $\lambda = -x_{\min}$ and the smallest $\lambda = -x_{\max}$. Neither of the intervals defined in the statement contain any roots and therefore we can conclude that $l(\lambda)$ will be continuous therein. Note that the function f_λ on $[x_{\min}, x_{\max}]$ is not even defined for λ in $(-x_{\max}, -x_{\min})$. □

Lemma 134 (Tightening with the matrix spectrum unknown). *Consider a finitely supported valid function a. Let $[x_{\min}(\gamma), x_{\max}(\gamma)]$ be the smallest interval containing supp(a_γ) (see Definition 109). Consider $m(\gamma, x_{\min}(\gamma), x_{\max}(\gamma))$ as a function of γ (see Definition 111). m has at least one root in the interval $(0, 1]$.*

Proof. To prove the claim it suffices to show that $l_\gamma(\lambda)$ has a root in the range $(0, \infty)$ for some $\gamma \in (0, 1]$. Note that we are given a valid function a which means $\text{supp}(a) \in \mathbb{R}_{\geq}$.

We assume that $l_{\gamma=1}(\lambda) > 0$ for all $\lambda \in (0, \infty)$ because if this was not the case then we trivially have $\gamma = 1$ as a root, i.e. $m(1, x_{\min}(1), x_{\max}(1)) = 0$.

Notice that since $\sum h(x) = \sum g(x)$ we have

$$\begin{aligned} \lambda l(\lambda) &= \sum h(x)(\lambda f_\lambda(x) + 1) - \sum g(x)(\lambda f_\lambda(x) + 1) \\ &= \sum h(x)\frac{x}{\lambda + x} - \sum g(x)\frac{x}{\lambda + x}. \end{aligned}$$

Therefore for the remainder of this proof we redefine $f_\lambda = \frac{1}{\lambda}\frac{x}{\lambda+x}$ without changing the value of l or by extension l_γ (the $1/\lambda$ factor is partly why we restricted λ to $(0, \infty)$ instead of the more general $(-x_{\min}, \infty)$). Note that $\lim_{\gamma \to 0^+} l_\gamma(\lambda) < 0$ for all $\lambda \in (0, \infty)$ because $h_\gamma(x) = h(x/\gamma)$ which means $\lim_{\gamma \to 0} \sum h_\gamma(x) f_\lambda(x) = \lim_{\gamma \to 0} \sum h(x) f_\lambda(\gamma x) = 0$ since $\lim_{x \to 0} f_\lambda(x) = 0$. This in turn means $\lim_{\gamma \to 0^+} l_\gamma(\lambda) = -\sum g(x) f_\lambda(x) < 0$.

Further, each term constituting $l_\gamma(\lambda)$ is finite for $\lambda \in (0, \infty)$ since for $\lambda > 0$ the denominators are of the form $\lambda + x$ which are always positive. Hence $l_\gamma(\lambda)$ as a function of $\lambda \in [0, \infty)$ and $\gamma \in (0, 1]$ is continuous. By continuity then between $\gamma = 0^+$ and $\gamma = 1$ there should be a root.

It remains to justify why we extended the range of λ from $(0, \infty)$ to $(-\infty, -x_{\max}) \cup (-x_{\min}, \infty)$ in the definition of m (see Definition 111) as it appears in the statement of the lemma. This is due to the fact that $l_\gamma(\lambda)$ is continuous for λ in the stated range, see Lemma 133, and so there might be a root which appears in the extended range. If this is the case we would like to use this possibly higher value of γ (because for a small enough value a non-negative root must appear due to the aforesaid reasoning). This can help us avoid infinities (we will explain this later). □

Once we are guaranteed that there is at least one perfect stretching amount, we want to know the spectrum of the matrices. We state a slightly more general result which is a direct consequence of the results from the previous subsubsection. The difference is that it is stated in a form that would be useful for the algorithm.

Lemma 135 (Matrix spectrum from a valid function)**.** *Consider a valid function a, i.e. an a such that $l(\lambda) \geq 0$ and $l^1 \geq 0$ for all $\lambda \in [0, \infty)$ (see Definition 109) and let $[\chi, \xi]$ be such that for all $\lambda \in [-\infty, -\xi) \cup (-\chi, \infty]$ we have $l(\lambda) \geq 0$.*

Then, there exists a legal COF, corresponding to the function a, with its spectrum contained in $[\chi, \xi]$.

Proof. Since $l(\lambda) \geq 0$ for $\lambda \in (-\infty, -\xi) \cup (-\chi, \infty)$ and $l^1 \geq 0$ we know from Corollary 123 that a is (χ, ξ) valid. From Lemma 129 we know that a is EBRM on $[\chi, \xi]$. Finally from Lemma 104 we know that there exists a legal COF with spectrum in $[\chi, \xi]$. □

Recall that in Subsection 1.3.3 we said that we focus on operator monotone functions f with the special property that f^{-1} is also an operator monotone. We now establish that f_λs (see Definition 107) have this property.

Lemma 136 ($H \geq G \iff f_\lambda(H) \geq f_\lambda(G)$)**.** *Let H, G be real symmetric matrices, $[\chi, \xi]$ be the smallest interval containing spec$[H \oplus G]$ and f_λ be on (χ, ξ) (see Definition 106; f_λ is defined for $\lambda \in \mathbb{R} \backslash [-\xi, -\chi]$). Then, $H \geq G$ if and only if $f_\lambda(H) \geq f_\lambda(G)$.*

Proof. $H \geq G \implies f_\lambda(H) \geq f_\lambda(G)$ because f_λ is an operator monotone function for matrices with spectrum in $[\chi, \xi]$. We prove the converse. We find the inverse function of f_λ and show that it is also an operator monotone. Start with recalling that for $x \in [\chi, \xi]$ we have

$$y = f_\lambda(x) = \frac{-1}{\lambda + x} \implies x = -\frac{1}{y} - \lambda$$

where $\lambda \in \mathbb{R} \backslash [\chi, \xi]$. Thus $f_\lambda^{-1}(y) = -\frac{1}{y} - \lambda$. For a given λ either $f_\lambda(\chi)$ and $f_\lambda(\xi)$ are both greater than zero or both less than zero. Hence the operator monotones $f'_{\lambda'}(y)$ on $[f_\lambda(\chi), f_\lambda(\xi)]$ permit $\lambda' = 0$. Consequently $f'_{\lambda'=0}(y) = \frac{-1}{y}$ is an operator monotone on $[f_\lambda(\chi), f_\lambda(\xi)]$. A constant is also an operator monotone. Thus we conclude $f_\lambda^{-1}(y)$ is an operator monotone on the required interval establishing the converse. □

This completes the second step of the groundwork.

6.3.2.3 For Wiggle-v; the infinite part

For the final step of the groundwork, we need to establish a result which lets us tackle the divergences head-on. What is this divergence issue? Recall form our discussion in Subsection 1.3.3, our strategy was to tighten (we saw hints of how that can be done in the previous sections) and then find the operator monotone f_λ for which the ellipsoids touch along the $|w\rangle$ direction. What happens if one

of the ellipsoids under the action of this operator monotone has infinite curvature along some directions? One can in fact show that there are cases where this would necessarily happen. Having infinite curvature means that the corresponding matrix has a divergence. It is like an ellipse gets mapped to a line. Our algorithm will fail in this situation because the normal at the tip of a line is ill defined.

Our strategy is to show that tightness is preserved under the action of f_λ. This means that if for some λ' we consider the ellipsoids obtained by applying $f_{\lambda'}$ and we find that they touch along $|w\rangle$ then for some other λ'' ($\neq \lambda'$) they would continue to touch but along some other direction. This allows us to consider the sequence leading to the divergence. We use this sequence in the analysis of the algorithm.

Here we start by showing this result in the case where everything is well defined and then extend it to the divergent case.

Lemma 137 (Strict inequality under f_λ)**.** *$H > G$ if and only if $f_\lambda(H) > f_\lambda(G)$ where f_λ is on $(\chi, \xi) \supset$ spec$[H \oplus G]$.*

Proof. Note that $H > G \iff H' := H + \lambda\mathbb{I} > G + \lambda\mathbb{I} =: G'$ where $\lambda \in \bar{\mathbb{R}}\backslash[-\xi, -\chi]$ (by definition of f_λ on (χ, ξ); see Definition 106). There can be two cases, either both the matrices are strictly positive or both are strictly negative. Let us assume the former (the other follows similarly). We have

$$\begin{aligned} & H' > G' > 0 \\ \iff & \mathbb{I} > H'^{-1/2} G' H'^{-1/2} \\ \iff & \mathbb{I} < H'^{1/2} G'^{-1} H'^{1/2} \\ \iff & H'^{-1} < G'^{-1} \end{aligned}$$

where the first inequality follows from the fact that multiplication by a positive matrix doesn't affect the inequality (hint: it only changes the vectors $|w\rangle$ we use to show $\langle w| (H' - G') |w\rangle > 0$ but maps the set of rays to themselves as the norm of the vector might change), the second follows from the fact that one can diagonalise the matrices (identity stays the same) and then it is just a set of inequalities involving real numbers, and the third follows from again multiplication by a positive matrix. The last one is the same as $f_\lambda(H) > f_\lambda(G)$. □

Corollary 138 (Tightness preservation under f_λ)**.** *Let $H \geq G$ and f_λ be on $(\chi, \xi) \supset$ spec$[H \oplus G]$. There exists a $|w\rangle$ such that $\langle w| (H - G) |w\rangle = 0$ if and only if there exists a $|w_\lambda\rangle$ such that $\langle w_\lambda| (f_\lambda(H) - f_\lambda(G)) |w_\lambda\rangle = 0$.*

Proof. The contrapositive of the aforesaid condition is that $f_\lambda(H) > f_\lambda(G)$ if and only if $H > G$ which holds due to Lemma 137. □

Lemma 139 (Extending tightness preservation under f_λ to apparently divergent situations)**.** *Let X_h, X_g be diagonal matrices with spec$[X_h] \in (\chi, \xi]$, spec$[X_g] \in [\chi, \xi)$ and let f_λ be on $[-\xi, -\chi]$. Let, further, O be an orthogonal matrix such that $X_h \geq O X_g O^T$.*

There exists a vector $|w\rangle$ such that $\langle w| \left(f_{-\xi}(X_h) - O f_{-\xi}(X_g) O^T\right) |w\rangle = 0$ if and only if there exists a $|w_\lambda\rangle$ such that $\langle w_\lambda| \left(f_\lambda(X_h) - O f_\lambda(X_g) O^T\right) |w_\lambda\rangle = 0$ for a $\lambda \in \mathbb{R}\backslash(-\xi, -\chi)$.

Similarly, there exists a vector $|w\rangle$ such that $\langle w| \left(f_{-\chi}(X_h) - O f_{-\chi}(X_g) O^T\right) |w\rangle = 0$ if and only if there exists a $|w_\lambda\rangle$ such that $\langle w_\lambda| \left(f_\lambda(X_h) - O f_\lambda(X_g) O^T\right) |w_\lambda\rangle = 0$ for a $\lambda \in \mathbb{R}\backslash(-\xi, -\chi)$.

Proof. The trouble with this version of the tightness statement is that X_h has an eigenvalue ξ (if it doesn't then it reduces to the previous statement) which means that $f_{-\xi}(X_h)$ is not well defined. We assume that X_h can be expressed as

$$X_h = \begin{bmatrix} X_h' & \\ & \xi\mathbb{I}'' \end{bmatrix}$$

where X'_h has no eigenvalue equal to ξ and $\mathbb{I}''$ is the identity matrix in the subspace. We can write

$$\begin{aligned}
&X_h > OX_gO^T \\
&\iff \begin{bmatrix} f_\lambda(X'_h) & \\ & f_\lambda(\xi\mathbb{I}'') \end{bmatrix} > Of_\lambda(X_g)O^T \text{ for } \lambda \in \mathbb{R}\backslash[-\xi,-\chi] \\
&\iff \begin{bmatrix} f_\lambda(X'_h) & \\ & \mathbb{I}'' \end{bmatrix} > \begin{bmatrix} \mathbb{I}' & \\ & f_\lambda(\xi\mathbb{I}'')^{-1/2} \end{bmatrix} Of_\lambda(X_g)O^T \begin{bmatrix} \mathbb{I}' & \\ & f_\lambda(\xi\mathbb{I}'')^{-1/2} \end{bmatrix} \text{ for } \lambda \in \mathbb{R}\backslash[-\xi,-\chi]
\end{aligned}$$

where in the last line the expression has a well defined limit for $\lambda = -\xi$. This establishes the contrapositive variant of the statement we wanted to prove (similar to the strategy used for proving Corollary 138) once we note the following. If $\langle w| \left(f_{-\xi}(X_h) - Of_{-\xi}(X_g)O^T\right)|w\rangle = 0$ it is easy to see that $\begin{bmatrix} 0 & \\ & \mathbb{I}'' \end{bmatrix}|w\rangle = 0$ otherwise due to the constraint on the spectrum of X_g the aforesaid expression would be ∞. This entails that

$$\langle w| \left(\begin{bmatrix} f_{-\xi}(X'_h) & \\ & \mathbb{I}'' \end{bmatrix} - \begin{bmatrix} \mathbb{I}' & \\ & f_{-\xi}(\xi\mathbb{I}'')^{-1/2} \end{bmatrix} Of_{-\xi}(X_g)O^T \begin{bmatrix} \mathbb{I}' & \\ & f_{-\xi}(\xi\mathbb{I}'')^{-1/2} \end{bmatrix} \right) |w\rangle = 0.$$

One can similarly prove the case for $f_{-\chi}(X_g)$. □

6.3.3 The Algorithm

We now describe our algorithm and formally state its correctness. Thereafter, we motivate each step of the algorithm and prove its correctness.

Definition 140 (EMA Algorithm)**.** Given a finitely supported function a (we assume it is Λ-valid (see Definition 62)) proceed in the following three phases.

PHASE 1: INITIALISATION

- **Tightening procedure**: Let $[x_{\min}(\gamma'), x_{\max}(\gamma')]$ be the support domain for $a_{\gamma'}$ (see Definition 109) where $\gamma' \in (0,1]$ is just a variable. Let $\gamma \in (0,1]$ be the largest root of $m(\gamma', x_{\min}(\gamma'), x_{\max}(\gamma'))$. Let $x_{\max} := x_{\max}(\gamma)$ and $x_{\min} := x_{\min}(\gamma)$.
- **Spectral domain for the representation**: Find the smallest interval $[\chi,\xi]$ such that $l_\gamma(\lambda) \geq 0$ for $\lambda \in \bar{\mathbb{R}}\backslash[\chi,\xi]$. If $\text{supp}(g)$, $\text{supp}(h)$ is not contained in $[\chi,\xi]$ then from all expansions of $[\chi,\xi]$ that contain the aforesaid sets, pick the smallest. Relabel this interval to $[\chi,\xi]$.
- **Shift**: Transform

  $$a(x) \to a'(x') := a(x' + \chi - 1)$$

  where instead of 1 any positive constant would do (justified by Corollary 128). Similarly transform

  $$\begin{aligned} g(x) &\to g'(x') := g(x' + \chi - 1) \\ h(x) &\to h'(x') := h(x' + \chi - 1). \end{aligned}$$

  Relabel a' to a, g' to g and h' to h. (Remark: We do not deduce h and g from a as its positive and negative part because they might now have common support due to the tightening procedure.)

- **The matrices**: For $n := n_g + n_h - 1$ we define $n \times n$ matrices with spectrum in $[\chi, \xi]$ and n dimensional vectors as

$$\begin{aligned} X_g^{(n)} &= \text{diag}[\chi, \chi, \dots x_{g_1}, x_{g_2} \dots, x_{g_{n_g}}], \\ X_{h_\gamma}^{(n)} &= \text{diag}[\gamma x_{h_1}, \gamma x_{h_2}, \dots, \gamma x_{h_{n_h}}, \xi, \xi, \dots], \\ \left|v^{(n)}\right\rangle &\doteq \left[0, 0 \dots, \sqrt{p_{g_1}}, \sqrt{p_{g_2}}, \dots, \sqrt{p_{g_{n_g}}}\right], \\ \left|w^{(n)}\right\rangle &\doteq \left[\sqrt{p_{h_1}}, \sqrt{p_{h_2}}, \dots, \sqrt{p_{h_{n_h}}}, 0, 0 \dots\right] \end{aligned}$$

  where $g = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ and $h = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$. Note that n_g and n_h may be different.

- **Bootstrapping the iteration**:
  - Basis: $\left\{\left|t_{h_i}^{(n+1)}\right\rangle\right\}$ where $\left|t_{h_i}^{(n+1)}\right\rangle := |i\rangle$ for $i = 1, 2 \dots n$ where $|i\rangle$ refers to the standard basis in which the matrices and the vectors were originally written.
  - Matrix Instance: $\underline{\mathrm{X}}^{(n)} = \{X_h^{(n)}, X_g^{(n)}, \left|w^{(n)}\right\rangle, \left|v^{(n)}\right\rangle\}$.

PHASE 2: ITERATION

- Objective: Find the objects $\left|u_h^{(k)}\right\rangle, \bar{O}_g^{(k)}, \bar{O}_h^{(k)}$ and $s^{(k)}$ (which together relate $O^{(k)}$ to $O^{(k-1)}$ where $O^{(k)}$ solves $\underline{\mathrm{X}}^{(k)}$ and $O^{(k-1)}$ solves $\underline{\mathrm{X}}^{(k-1)}$ that is yet to be defined)
- Input: We will assume we are given
  - Basis: $\left\{\left|t_{h_i}^{(k+1)}\right\rangle\right\}$
  - Matrix Instance: $\underline{\mathrm{X}}^{(k)} = \left(X_h^{(k)}, X_g^{(k)}, \left|w^{(k)}\right\rangle, \left|v^{(k)}\right\rangle\right)$ with attribute $\chi^{(k)} > 0$
  - Function Instance: $\underline{\mathrm{X}}^{(k)} \to \underline{\mathrm{x}}^{(k)} = \left(h^{(k)}, g^{(k)}, a^{(k)}\right)$
- Output:
  - Basis: $\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_i}^{(k)}\right\rangle\right\}$
  - Matrix Instance: $\underline{\mathrm{X}}^{(k-1)} = \left(X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle\right)$ with attribute $\chi^{(k-1)} > 0$
  - Function Instance: $\underline{\mathrm{X}}^{(k-1)} \to \underline{\mathrm{x}}^{(k-1)} = \left(h^{(k-1)}, g^{(k-1)}, a^{(k-1)}\right)$
  - Unitary Constructors: Either $\bar{O}_g^{(k)}$ and $\bar{O}_h^{(k)}$ are returned or $\bar{O}^{(k)}$ is returned. If $\bar{O}^{(k)}$ is returned, set $\bar{O}_g^{(k)} := \bar{O}^{(k)}$ and $\bar{O}_h^{(k)} = \mathbb{I}$.
  - Relation: If $s^{(k)}$ is not specified, define $s^{(k)} := 1$.
    If $s^{(k)} = 1$ then use

$$O^{(k)} := \bar{O}_h^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_g^{(k)}$$

    else use

$$O^{(k)} := \left[\bar{O}_h^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_g^{(k)}\right]^T.$$

- Algorithm:
  - **Boundary condition: If** $n_g = 0$ and $n_h = 0$ **then** set $k_0 = k$ and **jump to** phase 3.

– **Tighten**: Define $X^{(k)}_{h_{\gamma'}} := \gamma' X^{(k)}$ where $\gamma' \in (0,1]$ is just a variable. Let γ be the largest root of $m(\gamma', \chi^{(k)}_{\gamma'}, \xi^{(k)}_{\gamma'})$ for $a^{(k)}$ where $\chi^{(k)}_{\gamma'}, \xi^{(k)}_{\gamma'}$ are such that $[\chi^{(k)}_{\gamma'}, \xi^{(k)}_{\gamma'}]$ is the smallest interval containing $\mathrm{spec}[X^{(k)}_{h_{\gamma'}} \oplus X^{(k)}_g]$. Relabel $X^{(k)}_{h_\gamma}$ to $X^{(k)}_h$, $\chi^{(k)}_\gamma$ to $\chi^{(k)}$ and $\xi^{(k)}_\gamma$ to $\xi^{(k)}$ for notational ease. Similarly relabel $a^{(k)}_\gamma$ to $a^{(k)}$, $h^{(k)}_\gamma$ to $h^{(k)}$, $l^{(k)}_\gamma$ to $l^{(k)}$. Update $x_{\min}$ and $x_{\max}$ to be such that $\mathrm{supp}(a^{(k)}) \in [x^{(k)}_{\min}, x^{(k)}_{\max}]$ is the smallest such interval. Define $s^{(k)} := 1$.

– **Honest align**: **If** $l^{1(k)} = 0$ **then** define $\eta = -\chi^{(k)} + 1$

$$X'^{(k)}_h := X^{(k)}_h + \eta, \quad X'^{(k)}_g := X_g + \eta.$$

Else: Pick a root λ of the function $l^{(k)}(\lambda')$ in the domain $\mathbb{R}\backslash(-\xi^{(k)}, -\chi^{(k)})$. In the following two cases we consider the function f_λ on $[\chi^{(k)}, \xi^{(k)}]$.

* If $\lambda \neq -\chi^{(k)}$ then: Let $\eta = -f_\lambda(\chi^{(k)}) + 1$ where any positive constant could be chosen instead of 1. Define

$$X'^{(k)}_h := f_\lambda(X^{(k)}_h) + \eta, \quad X'^{(k)}_g := f_\lambda(X^{(k)}_g) + \eta.$$

* If $\lambda = -\chi^{(k)}$ then: Update $s^{(k)} = -1$. Let $\eta = -f_\lambda(\xi^{(k)}) - 1$ where any positive constant could be chosen instead of 1. Define

$$X'^{(k)}_h := X''^{(k)}_g, \quad X'^{(k)}_g := X''^{(k)}_h,$$

where

$$X''^{(k)}_h := -f_\lambda(X^{(k)}_h) - \eta, \quad X''^{(k)}_g := -f_\lambda(X^{(k)}_g) - \eta$$

and make the replacement

$$\left|v^{(k)}\right\rangle \to \left|w^{(k)}\right\rangle$$
$$\left|w^{(k)}\right\rangle \to \left|v^{(k)}\right\rangle.$$

– **Remove spectral collision: If** $\lambda = -\chi^{(k)}$ or $\lambda = -\xi^{(k)}$ **then**

1. **Idle point: If** for some j', j, we have $q^{(k)}_{g_{j'}} = q^{(k)}_{h_j}$ and $y^{(k)}_{g_{j'}} = y^{(k)}_{h_j}$ **then** the solution is given by Definition 142
Jump to **End**.
2. **Final Extra**: **If** for some j, j' we have $q^{(k)}_{g_{j'}} > q^{(k)}_{h_j}$ and $y^{(k)}_{g_{j'}} = y^{(k)}_{h_j}$ **then** the solution is given by Definition 143
Jump to **End**.
3. **Initial Extra**: **If** for some j, j' we have $q^{(k)}_{g_{j'}} < q^{(k)}_{h_j}$ and $y^{(k)}_{g_{j'}} = y^{(k)}_{h_j}$ **then** the solution is given by Definition 144
Jump to **End**.

– **Evaluate the Reverse Weingarten Map**:

1. Consider the point $\left|w^{(k)}\right\rangle / \sqrt{\left\langle w^{(k)}\right| X'^{(k)}_h \left|w^{(k)}\right\rangle}$ on the ellipsoid $X'^{(k)}_h$. Evaluate the normal at this point as $\left|u^{(k)}_h\right\rangle = \mathcal{N}\left(\sum_{i=1}^{n^{(k)}_h} \sqrt{p^{(k)}_{h_i}} x'^{(k)}_{h_i} \left|t^{(k+1)}_{h_i}\right\rangle\right)$. Similarly evaluate $\left|u^{(k)}_g\right\rangle$, the normal at the point $\left|v^{(k)}\right\rangle / \sqrt{\left\langle w^{(k)}\right| X'^{(k)}_g \left|w^{(k)}\right\rangle}$ on the ellipsoid $X'^{(k)}_g$.

2. Recall that for a given diagonal matrix $X = \sum_i y_i |i\rangle \langle i| > 0$ and normal vector $|u\rangle = \sum_i u_i |i\rangle$ the Reverse Weingarten map is given by $W_{ij} = \left(-\frac{y_j^{-1} y_i^{-1} u_i u_j}{r^2} + y_i^{-1} \delta_{ij} \right)$ where $r = \sqrt{\sum y_i^{-1} u_i^2}$. Evaluate the Reverse Weingarten maps $W_h^{\prime(k)}$ and $W_g^{\prime(k)}$ along $\left|u_h^{(k)}\right\rangle$ and $\left|u_g^{(k)}\right\rangle$ respectively.

3. Find the eigenvectors and eigenvalues of the Reverse Weingarten maps. The eigenvectors of W_h' form the h tangent (and normal) vectors $\left\{ \left\{ \left|t_{h_i}^{(k)}\right\rangle \right\}, \left|u_h^{(k)}\right\rangle \right\}$. The corresponding radii of curvature are obtained from the eigenvalues $\left\{ \{r_{h_i}^{(k)}\}, 0 \right\} = \left\{ \{c_{h_i}^{(k)-1}\}, 0 \right\}$ which are inverses of the curvature values. The tangents are labelled in the decreasing order of radii of curvature (increasing order of curvature). Similarly for the g tangent (and normal) vectors. Fix the sign freedom in the eigenvectors by requiring $\left\langle t_{h_i}^{(k)} | w^{(k)} \right\rangle \geq 0$ and $\left\langle t_{g_i}^{(k)} | v^{(k)} \right\rangle \geq 0$.

– **Finite Method: If** $\lambda \neq -\xi^{(k)}$ and $\lambda \neq -\chi^{(k)}$, i.e. if it is the finite case **then**

1. $\bar{O}^{(k)} := \left|u_h^{(k)}\right\rangle \left\langle u_g^{(k)}\right| + \sum_{i=1}^{k-1} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{g_i}^{(k)}\right|$
2. $\left|v^{(k-1)}\right\rangle := \bar{O}^{(k)} \left|v^{(k)}\right\rangle - \left\langle u_h^{(k)}\right| \bar{O}^{(k)} \left|v^{(k)}\right\rangle \left|u_h^{(k)}\right\rangle$ and $\left|w^{(k-1)}\right\rangle := \left|w^{(k)}\right\rangle - \left\langle u_h^{(k)} | w^{(k)} \right\rangle \left|u_h^{(k)}\right\rangle$.
3. Define $X_h^{(k-1)} := \mathrm{diag}\{c_{h_1}^{(k)}, c_{h_2}^{(k)} \ldots, c_{h_{k-1}}^{(k)}\}$, $X_g^{(k-1)} := \mathrm{diag}\{c_{g_1}^{(k)}, c_{g_2}^{(k)} \ldots c_{g_{k-1}}^{(k)}\}$.
4. **Jump** to **End**.

– **Wiggle-v Method: If** $\lambda = -\xi^{(k)}$ or $\lambda = -\chi^{(k)}$ **then**

1. $\left|u_h^{(k)}\right\rangle$ is renamed to $\left|\bar{u}_h^{(k)}\right\rangle$, $\left|u_g^{(k)}\right\rangle$ remains the same.
2. Let $\tau = \cos\theta := \left\langle u_g^{(k)} | v^{(k)} \right\rangle / \left\langle \bar{u}_h^{(k)} | w^{(k)} \right\rangle$. Let $\left|\bar{t}_h^{(k)}\right\rangle$ be an eigenvector of $X_h^{\prime(k)-1}$ with zero eigenvalue (comment: this is also perpendicular to $\left|w^{(k)}\right\rangle$). Redefine

$$\left|u_h^{(k)}\right\rangle := \cos\theta \left|\bar{u}_h^{(k)}\right\rangle + \sin\theta \left|\bar{t}_h^{(k)}\right\rangle,$$

$$\left|t_{h_k}^{(k)}\right\rangle = s \left(-\sin\theta \left|\bar{u}_h^{(k)}\right\rangle + \cos\theta \left|\bar{t}_h^{(k)}\right\rangle \right)$$

where the sign $s \in \{1, -1\}$ is fixed by demanding $\left\langle t_{h_k}^{(k)} | w^{(k)} \right\rangle \geq 0$.

3. $\bar{O}^{(k)}$ and $\left|v^{(k-1)}\right\rangle, \left|w^{(k-1)}\right\rangle$ are evaluated as step i and ii of the finite case (a).
4. Define

$$X_h^{\prime(k-1)} := \mathrm{diag}\{c_{h_1}^{(k)}, c_{h_2}^{(k)}, \ldots, c_{h_{k-1}}^{(k)}\}, \qquad X_g^{\prime(k-1)} := \mathrm{diag}\{c_{g_1}^{(k)}, c_{g_2}^{(k)}, \ldots, c_{g_{k-1}}^{(k)}\}.$$

Let $[\chi^{\prime(k-1)}, \xi^{\prime(k-1)}]$ denote the smallest interval containing $\mathrm{spec}[X_h^{\prime(k-1)} \oplus X_g^{\prime(k-1)}]$. Let $\lambda' = -\chi^{\prime(k-1)} + 1$ where instead of 1 any positive number would also work. Consider $f_{\lambda''}$ on $[\chi^{\prime(k-1)}, \xi^{\prime(k-1)}]$. Let $\eta = -f_{\lambda'}(\chi^{\prime(k-1)}) + 1$. Define

$$X_h^{(k-1)} := f_{\lambda'}(X_h^{\prime(k-1)}) + \eta, \qquad X_g^{(k-1)} := f_{\lambda'}(X_g^{\prime(k-1)}) + \eta.$$

5. **Jump** to **End**.

– **End**: Restart the current phase (phase 2) with the newly obtained $(k-1)$ sized objects.

PHASE 3: RECONSTRUCTION

Let k_0 be the iteration at which the algorithm stops. Using the relation

$$O^{(k)} = \bar{O}_g^{(k)} \left(\left| u_h^{(k)} \right\rangle \left\langle u_h^{(k)} \right| + O^{(k-1)} \right) \bar{O}_h^{(k)}$$

(or its transpose if $s^{(k)} = -1$), evaluate $O^{(k_1)}$ from $O^{(k_0)} := \mathbb{I}_{k_0}$, then $O^{(k_2)}$ from $O^{(k_1)}$, then $O^{(k_3)}$ from $O^{(k_2)}$ and so on until $O^{(n)}$ is obtained which solves the matrix instance $\underline{X}^{(n)}$ we started with. In terms of EBRM matrices, the solution is given by $H = X_h^{(n)}$, $G = O^{(n)} X_g O^{(n)T}$, and $|w\rangle = \left|w^{(n)}\right\rangle$.

Theorem 141 (Correctness of the EMA Algorithm)**.** *Given a Λ-valid function, the EMA algorithm (see Definition 140) always finds an orthogonal matrix O of size at most $n \times n$ where $n = n_g + n_h$, such that the constraints on O stated in Theorem 10 corresponding to the function a, are satisfied.*

Definition 142 (Spectral Collision: Case Idle Point)**.**

$$\left\{ \left| u_h^{(k)} \right\rangle, \left| t_{h_1}^{(k)} \right\rangle, \left| t_{h_2}^{(k)} \right\rangle, \dots \left| t_{h_{k-1}}^{(k)} \right\rangle \right\} \overset{\text{componentwise}}{:=}$$
$$\left\{ \left| t_{h_j}^{(k+1)} \right\rangle, \left| t_{h_1}^{(k+1)} \right\rangle, \left| t_{h_2}^{(k+1)} \right\rangle, \dots \left| t_{h_{j-1}}^{(k+1)} \right\rangle, \left| t_{h_{j+1}}^{(k+1)} \right\rangle, \dots \left| t_{h_k}^{(k+1)} \right\rangle \right\},$$

$$\bar{O}^{(k)} := \sum_{i=1}^{k} |a_i\rangle \left\langle t_{h_i}^{(k+1)} \right|,$$

where

$$\{|a_1\rangle, |a_2\rangle \dots |a_k\rangle\} \overset{\text{componentwise}}{:=}$$

$$\begin{cases} \left\{ \left| t_{h_1}^{(k+1)} \right\rangle, \left| t_{h_2}^{(k+1)} \right\rangle, \dots \left| t_{h_{j-1}}^{(k+1)} \right\rangle, \left| t_{h_{j'}}^{(k+1)} \right\rangle, \left| t_{h_j}^{(k+1)} \right\rangle, \left| t_{h_{j+1}}^{(k+1)} \right\rangle, \right. \\ \qquad \left. \dots, \left| t_{h_{j'-1}}^{(k+1)} \right\rangle, \left| t_{h_{j'+1}}^{(k+1)} \right\rangle \dots \left| t_{h_k}^{(k+1)} \right\rangle \right\} & j < j' \\ \left\{ \left| t_{h_1}^{(k+1)} \right\rangle, \left| t_{h_2}^{(k+1)} \right\rangle, \dots \left| t_{h_{j'-1}}^{(k+1)} \right\rangle, \left| t_{h_{j'+1}}^{(k+1)} \right\rangle \dots \right. \\ \left. \left| t_{h_{j-1}}^{(k+1)} \right\rangle, \left| t_{h_{j'}}^{(k+1)} \right\rangle, \left| t_{h_j}^{(k+1)} \right\rangle, \left| t_{h_{j+1}}^{(k+1)} \right\rangle \dots \left| t_{h_k}^{(k+1)} \right\rangle \right\} & j > j' \\ \left\{ \left| t_{h_1}^{(k+1)} \right\rangle, \left| t_{h_2}^{(k+1)} \right\rangle, \dots \left| t_{h_k}^{(k+1)} \right\rangle \right\} & j = j', \end{cases}$$

and

$$X_h^{(k-1)} := \sum_{i \neq j} y_{h_i}^{(k)} \left| t_{h_i}^{(k+1)} \right\rangle \left\langle t_{h_i}^{(k+1)} \right|,$$

$$X_g^{(k-1)} := \bar{O}^{(k)} X_g^{(k)} \bar{O}^{(k)T} - y_{h_j} \left| t_{h_j}^{(k+1)} \right\rangle \left\langle t_{h_j}^{(k+1)} \right|,$$

$$\left| w^{(k-1)} \right\rangle = \mathcal{N} \left[\left| w^{(k)} \right\rangle - \sqrt{p_{h_j}} \left| t_{h_j}^{(k+1)} \right\rangle \right], \quad \left| v^{(k-1)} \right\rangle = \mathcal{N} \left[\bar{O}^{(k)} \left| v^{(k)} \right\rangle - \sqrt{p_{h_j}} \left| t_{h_j}^{(k+1)} \right\rangle \right].$$

(This specifies $\underline{X}^{(k-1)} := \{X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle\}$.)

Definition 143 (Spectral Collision: Case Final Extra)**.**

$\underline{X}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|, X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|, \left|v^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{g_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right], \left|w^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{h_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$ where the coordinates and weights are given by

$$\begin{aligned}
\left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)}, q_{h_2}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_{j+1}}^{(k)}, \ldots q_{h_k}^{(k)}\right\} \\
\left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}, q_{g_{j'+1}}^{(k)}, q_{g_{j'+2}}^{(k)} \ldots q_{g_k}^{(k)}\right\} \\
\left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{g_2}^{(k)}, \ldots y_{g_k}^{(k)}\right\} \\
\left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{j-1}}^{(k)}, y_{h_{j+1}}^{(k)} \ldots, y_{h_k}^{(k)}\right\},
\end{aligned}$$

the basis is given by

$$\begin{aligned}
&\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=} \\
&\left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}.
\end{aligned}$$

The orthogonal matrices are given by $\bar{O}_h^{(k)} := \sum \left|t_{h_i}^{(k+1)}\right\rangle \langle a_i|$ where

$$\{|a_1\rangle, \ldots |a_k\rangle\} \to \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\},$$

$\bar{O}_g^{(k)} := \tilde{O}^{(k)} \bar{O}_h^{(k)}$ where

$$\begin{aligned}
\tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle + \sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}} \left|t_{h_{j'}}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_1}^{(k)}} \left\langle u_h^{(k)}\right| + \sqrt{q_{g_{j'}}^{(k)}} \left\langle t_{h_{j'}}^{(k)}\right|\right] \\
&+ \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle - \sqrt{q_{h_j}^{(k)}} \left|t_{h_{j'}}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left\langle u_h^{(k)}\right| - \sqrt{q_{g_1}^{(k)}} \left\langle t_{h_{j'}}^{(k)}\right|\right] \\
&+ \sum_{i \in \{1, \ldots k\} \setminus j'} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|.
\end{aligned}$$

Definition 144 (Spectral Collision: Case Initial Extra)**.**

$\underline{X}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|, X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|, \left|v^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{g_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right], \left|w^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{h_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$ where the coordinates and weights are given by

$$\begin{aligned}
\left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}, q_{h_{j+1}}^{(k)}, q_{h_{j+2}}^{(k)} \ldots q_{h_{k-1}}^{(k)}\right\} \\
\left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{g_1}^{(k)}, q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'+1}}^{(k)}, \ldots q_{g_k}^{(k)}\right\} \\
\left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{g_1}^{(k)}, \ldots y_{g_{j'-1}}^{(k)}, y_{g_{j'+1}}^{(k)} \ldots, y_{g_k}^{(k)}\right\} \\
\left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{k-1}}^{(k)}\right\},
\end{aligned}$$

the basis is given by

$$\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \dots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=}$$
$$\left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \dots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \dots \left|t_{h_k}^{(k+1)}\right\rangle\right\}.$$

The orthogonal matrices are given by $\bar{O}_h^{(k)} := \tilde{O}^{(k)} \sum |a_i\rangle \left\langle t_{h_i}^{(k+1)}\right|$ where

$$\{|a_1\rangle, \dots |a_k\rangle\} \overset{\text{componentwise}}{=} \left\{\left|t_{h_1}^{(k)}\right\rangle, \left|t_{h_2}^{(k)}\right\rangle \dots \left|t_{h_{k-1}}^{(k)}\right\rangle, \left|u_h^{(k)}\right\rangle\right\}.$$

$$\begin{aligned}\tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}}\left|u_h^{(k)}\right\rangle + \sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}}\left|t_{h_j}^{(k)}\right\rangle\right]\mathcal{N}\left[\sqrt{q_{h_k}^{(k)}}\left\langle u_h^{(k)}\right| + \sqrt{q_{g_j}^{(k)}}\left\langle t_{h_j}^{(k)}\right|\right] \\ &+ \mathcal{N}\left[\sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}}\left|u_h^{(k)}\right\rangle - \sqrt{q_{g_{j'}}^{(k)}}\left|t_{h_j}^{(k)}\right\rangle\right]\mathcal{N}\left[\sqrt{q_{g_j}^{(k)}}\left\langle u_h^{(k)}\right| - \sqrt{q_{h_k}^{(k)}}\left\langle t_{h_j}^{(k)}\right|\right] \\ &+ \sum_{i\in\{1,\dots k\}\setminus j}\left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|\end{aligned}$$

and $\bar{O}_h^{(k)}$ is given by the basis change $\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \dots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \to \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \dots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\}$.

We start with motivating the exact step of the algorithm and then provide a proof or justification for the claims made in that step.

6.3.3.1 Phase 1: Initialisation

We are given a Λ-valid transition $g \to h$ and the EBRM function $a = h - g$. (Remark: We use below the notation used in the definition of a transition.)

Since the function is EBRM we know there are matrices $H \geq G$ and a vector $|\psi\rangle$ such that $a = \text{Prob}[H, |\psi\rangle] - \text{Prob}[G, |\psi\rangle]$. We also know that the maximum matrix size we need to consider is $n_g + n_h - 1$. We want to know the spectrum of the matrices involved to proceed.

The picture we have in mind is the following. We know that $H \geq G$ in terms of ellipsoids means that the H ellipsoid is inside the G ellipsoid (the order gets reversed). We try to expand the H ellipsoid (which means scaling down the matrix H) until it touches the G ellipsoid. When they touch we know that the corresponding spectrum of the matrices is optimal in some sense. This would be trivial if we already knew H and G but it serves as a good picture nonetheless.

What we do know is the function $a = h - g$. We use the equivalence between EBRM and valid functions to perform the aforesaid tightening procedure even without knowing the matrices. We use $a_\gamma = h_\gamma - g$ where $h_\gamma(x) = h(x/\gamma)$ and check if a_γ stays valid as we shrink γ from one to zero. We stop the moment we see any signature of tightness. Using this a_γ we determine the spectrum of the matrices certifying the EBRM claim.

We start with tightening till we find some operator monotone labelled by λ for which $l_{\gamma'}(\lambda)$ disappears. This captures the notion of the ellipsoids touching as after applying this operator monotone, along the $|w\rangle$ direction, the ellipsoids must touch.

Tightening procedure: Let $[x_{\min}(\gamma'), x_{\max}(\gamma')]$ be the support domain for $a_{\gamma'}$ where $\gamma' \in (0, 1]$ is just a variable. Let $\gamma \in (0, 1]$ be the largest root of $m(\gamma', x_{\min}(\gamma'), x_{\max}(\gamma'))$. Let $x_{\max} := x_{\max}(\gamma)$ and $x_{\min} := x_{\min}(\gamma)$.

First we must show that there would indeed be a root of m as a function of γ' in the range $(0, 1]$. This is a direct consequence of Lemma 134. Second we must show that if we can find the matrices certifying a_γ is EBRM we can find the matrices certifying a is EBRM. This follows from the observation that $\gamma X_h \geq O X_g O^T$ implies that $X_h \geq \gamma X_h \geq O X_g O^T$.

We found a signature of tightness. Now we find the spectrum of the matrices involved.

Spectral domain for the representation: Find the smallest interval $[\chi, \xi]$ such that $l_\gamma(\lambda) \geq 0$ for $\lambda \in \bar{\mathbb{R}} \backslash [\chi, \xi]$. If $\text{supp}(g), \text{supp}(h)$ is not contained in $[\chi, \xi]$ then from all expansions of $[\chi, \xi]$ that contain the aforesaid sets, pick the smallest. Relabel this interval to $[\chi, \xi]$.

The interval so obtained will contain the spectrum of the matrices that certify a_γ is EBRM. This is justified by Lemma 135 using the fact that $l^1_\gamma \geq 0$ due to the previous step.

We need our matrices to be positive to be able to use the elliptic picture. We therefore shift the spectrum of the matrices so that the smallest eigenvalue required is one (where we could have used any positive number).

Shift: Transform

$$a(x) \to a'(x') := a(x' + \chi - 1)$$

where instead of 1 any positive constant would do (justified by Corollary 128). Similarly transform

$$\begin{aligned} g(x) &\to g'(x') := g(x' + \chi - 1) \\ h(x) &\to h'(x') := h(x' + \chi - 1). \end{aligned}$$

Relabel a' to be a, g' to be g and h' to be h. (Remark: We do not deduce h and g from a as its positive and negative part because they might now have common support due to the tightening procedure.)

We use Corollary 128 to deduce that if $a(x)$ is EBRM with spectrum in $[\chi, \xi]$ then $a'(x') = a(x' + \chi - 1)$ is EBRM with spectrum in $[1, \xi - \chi + 1]$. We must also show that if we can find the matrices certifying a' is EBRM then we can find the matrices certifying a is EBRM. This is a direct consequence of the fact that $X'_h \geq O X'_g O^T \iff X_h - (\chi - 1)\mathbb{I} \geq O(X_g - (\chi - 1)\mathbb{I})O^T$. The orthogonal matrix, O, which is of primary interest remains unchanged.

With the spectrum determined and adjusted to the elliptic picture, which we put to use soon, we fix everything except the orthogonal matrix by using the Canonical Orthogonal Form (up to a permutation).

The matrices: For $n := n_g + n_h - 1$ we define $n \times n$ matrices with spectrum in $[\chi, \xi]$ and n dimensional vectors as

$$\begin{aligned} X_g^{(n)} &= \text{diag}[\chi, \chi, \dots x_{g_1}, x_{g_2} \dots, x_{g_{n_g}}], \\ X_{h_\gamma}^{(n)} &= \text{diag}[\gamma x_{h_1}, \gamma x_{h_2}, \dots, \gamma x_{h_{n_h}}, \xi, \xi, \dots], \\ \left| v^{(n)} \right\rangle &\doteq \left[0, 0 \dots, \sqrt{p_{g_1}}, \sqrt{p_{g_2}}, \dots, \sqrt{p_{g_{n_g}}} \right], \\ \left| w^{(n)} \right\rangle &\doteq \left[\sqrt{p_{h_1}}, \sqrt{p_{h_2}}, \dots, \sqrt{p_{h_{n_h}}}, 0, 0 \dots \right] \end{aligned}$$

where $g = \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ and $h = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$. Note that n_g and n_h may be different.

We use Lemma 130 to deduce that $g \to h$ is a valid transition from the validity of a. Then we use Lemma 104 to write the diagonal matrices as described above given the valid transition $g \to h$, upto

a permutation. Our objective is to find a matrix $O^{(n)}$ such that $O^{(n)} \left|v^{(n)}\right\rangle = \left|w^{(n)}\right\rangle$ while satisfying the inequality $X_h^{(n)} \geq O^{(n)} X_g^{(n)} O^{(n)T}$.

We now remove all the redundant information and pack it into a form which we can iteratively reduce to a simpler form.

Bootstrapping the iteration:

- Basis: $\left\{\left|t_{h_i}^{(n+1)}\right\rangle\right\}$ where $\left|t_{h_i}^{(n+1)}\right\rangle := |i\rangle$ for $i = 1, 2 \ldots n$ where $|i\rangle$ refers to the standard basis in which the matrices and the vectors were originally written.
- Matrix Instance: $\underline{\mathrm{X}}^{(n)} = \{X_h^{(n)}, X_g^{(n)}, \left|w^{(n)}\right\rangle, \left|v^{(n)}\right\rangle\}$.

6.3.3.2 Phase 2: Iteration

We take as input the matrices X_g, X_h together with the vectors $|w\rangle, |v\rangle$ and churn out the same objects with one less dimension. We also output objects that, once we have iteratively reduced the problem to triviality, can be put together to find the orthogonal matrix O. See Figure 6.2 for a schematic reference.

- Objective: Find the objects $\left|u_h^{(k)}\right\rangle, \bar{O}_g^{(k)}, \bar{O}_h^{(k)}$ and $s^{(k)}$ (which together relate $O^{(k)}$ to $O^{(k-1)}$ where $O^{(k)}$ solves $\underline{\mathrm{X}}^{(k)}$ and $O^{(k-1)}$ solves $\underline{\mathrm{X}}^{(k-1)}$ that is yet to be defined)
- Input: We assume we are given
  - Basis: $\left\{\left|t_{h_i}^{(k+1)}\right\rangle\right\}$
  - Matrix Instance: $\underline{\mathrm{X}}^{(k)} = \left(X_h^{(k)}, X_g^{(k)}, \left|w^{(k)}\right\rangle, \left|v^{(k)}\right\rangle\right)$ with attribute $\chi^{(k)} > 0$
  - Function Instance: $\underline{\mathrm{X}}^{(k)} \to \underline{\mathrm{x}}^{(k)} = \left(h^{(k)}, g^{(k)}, a^{(k)}\right)$
- Output:
  - Basis: $\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_i}^{(k)}\right\rangle\right\}$
  - Matrix Instance: $\underline{\mathrm{X}}^{(k-1)} = \left(X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle\right)$ with attribute $\chi^{(k-1)} > 0$
  - Function Instance: $\underline{\mathrm{X}}^{(k-1)} \to \underline{\mathrm{x}}^{(k-1)} = \left(h^{(k-1)}, g^{(k-1)}, a^{(k-1)}\right)$
  - Unitary Constructors: Either $\bar{O}_g^{(k)}$ and $\bar{O}_h^{(k)}$ are returned or $\bar{O}^{(k)}$ is returned. If $\bar{O}^{(k)}$ is returned, set $\bar{O}_g^{(k)} := \bar{O}^{(k)}$ and $\bar{O}_h^{(k)} = \mathbb{I}$.
  - Relation: If $s^{(k)}$ is not specified, define $s^{(k)} := 1$.
    If $s^{(k)} = 1$ then use

$$O^{(k)} := \bar{O}_h^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_g^{(k)}$$

    else use

$$O^{(k)} := \left[\bar{O}_h^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_g^{(k)}\right]^T.$$

Our task is to solve the matrix instance $\underline{\mathrm{X}}^{(k)}$, i.e. find a real unitary $O^{(k)}$ such that

$X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T}$ and $O^{(k)} \left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle$. We assume that the solution exists and show that the solution to the smaller instance, denoted by $\underline{X}^{(k-1)}$ must also exist. More precisely, we show that $O^{(k)}$ must have the form $O^{(k)} = \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}^{(k)}$ (for a solution to exist) which satisfies the aforesaid constraints granted we can find $O^{(k-1)}$ which acts on a $k-1$ dimensional Hilbert space orthogonal to $\left|u_h^{(k)}\right\rangle$ and satisfies constraints of the same form in the smaller dimension, viz. $X_h^{(k-1)} \geq O^{(k-1)} X_g^{(k-1)} O^{(k-1)T}$ and $O^{(k-1)} \left|v^{(k-1)}\right\rangle = \left|w^{(k-1)}\right\rangle$. Hence the assumption that $O^{(k)}$ has a solution allows us to deduce that $O^{(k-1)}$ must also have a solution. This allow us to iteratively solve the problem.

In certain trivial cases, where a point appears both before and after a transition viz. $X_g^{(k)}$ and $X_h^{(k)}$ have a common eigenvalue, the solution has the form

$$O^{(k)} = \bar{O}_h^{(k)} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}_g^{(k)}.$$

Finally, in one of the "infinite" cases denoted by the "Wiggle-v method" the solution will have the form

$$O^{(k)} = \left[\left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}^{(k)}\right]^T.$$

- Algorithm:

If we reach a stage where the vector constraints have disappeared then we can simply stop.

  - **Boundary condition: If** $n_g = 0$ and $n_h = 0$ **then** set $k_0 = k$ and **jump to** phase 3.

    We assumed that an $O^{(k)}$ satisfying the constraints (listed right after the input/output section) exists. In this case it means that there exists an $O^{(k)}$ such that $X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T}$ as there is no vector $\left|v^{(k)}\right\rangle, \left|w^{(k)}\right\rangle$ to impose further constraints. Using Corollary 132 with $H = X_h^{(k)}$ and $G = O^{(k)} X_g^{(k)} O^{(k)T}$ we conclude that $O^{(k)}$ need only be a permutation matrix. Note that this step can never be entered right after the $\underline{X}^{(n)}$ instance as we start with assuming $n_g, n_h > 0$. Further, since the protocol by construction always returns X_h and X_g in the ascending order the permutation matrix will be $\mathbb{I}$.

Finally, we deal with the interesting case. We again use the picture where the H ellipsoid is contained inside the G ellipsoid. We expand the H ellipsoid (which corresponds to shrinking the H matrix) until it touches the G ellipsoid as before by working with the function a.

  - **Tighten**: Define $X_{h_{\gamma'}}^{(k)} := \gamma' X^{(k)}$ where $\gamma' \in (0, 1]$ is a variable. Let γ be the largest root of $m(\gamma', \chi_{\gamma'}^{(k)}, \xi_{\gamma'}^{(k)})$ for $a^{(k)}$ where $\chi_{\gamma'}^{(k)}, \xi_{\gamma'}^{(k)}$ are such that $[\chi_{\gamma'}^{(k)}, \xi_{\gamma'}^{(k)}]$ is the smallest interval containing $\mathrm{spec}[X_{h_{\gamma'}}^{(k)} \oplus X_g^{(k)}]$. Relabel $X_{h_\gamma}^{(k)}$ to $X_h^{(k)}$, $\chi_\gamma^{(k)}$ to $\chi^{(k)}$ and $\xi_\gamma^{(k)}$ to $\xi^{(k)}$ for notational ease. Similarly relabel $a_\gamma^{(k)}$ to $a^{(k)}$, $h_\gamma^{(k)}$ to $h^{(k)}$, $l_\gamma^{(k)}$ to $l^{(k)}$. Update $x_{\min}$ and $x_{\max}$ to be such that $\mathrm{supp}(a^{(k)}) \in [x_{\min}^{(k)}, x_{\max}^{(k)}]$ is the smallest such interval. Define $s^{(k)} := 1$.

Our burden again is to show that m as a function of γ' has a root. Unlike the first tightening procedure this time we know the spectrum of the matrices involved. Since we are given (by assumption) that the matrix instance has a solution we know that $l_{\gamma'=1}(\lambda) \geq 0$ and $l^1_{\gamma'=1} \geq 0$ for $\lambda \in \bar{\mathbb{R}} \backslash [\chi^{(k)}_{\gamma'=1}, \xi^{(k)}_{\gamma'=1}]$ using Lemma 129. We also know that $\chi^{(k)}_{\gamma'} > 0$ which means that $a^{(k)}$ (as deduced from the function instance of $\underline{\mathrm{X}}^{(k)}$) is a valid function. This observation lets us conclude that m as a function of γ' has a root in the required range because the reasoning behind a similar claim proved in Lemma 134 goes through unchanged.

The tightening procedure guarantees we will be able to find a λ which corresponds to an operator monotone such that after applying this function the ellipsoids, which we do not even know completely yet, must touch along the $|w\rangle$ direction. This piece of information is key to reducing the problem to a smaller instance of itself. Recall the picture with the H ellipsoid contained inside the G ellipsoid. If we know that they, in addition, touch at some known point then it must be so that the inner ellipsoid is more curved than the outer ellipsoid. When expressed algebraically, this condition essentially becomes that requirement that an ellipsoid $H^{(k-1)}$ that encodes the curvature of the ellipsoid $H^{(k)}$ at the point of contact must be contained inside the corresponding $G^{(k-1)}$ ellipsoid which encodes the curvature of the $G^{(k)}$ ellipsoid. The vector condition also reduces similarly. Subtleties arise when λ happens to have boundary values in its allowed range as this yields infinities and this has an interesting consequence.

– **Honest align**: **If** $l^{1(k)} = 0$ **then** define $\eta = -\chi^{(k)} + 1$

$$X_h'^{(k)} := X_h^{(k)} + \eta, \quad X_g'^{(k)} := X_g + \eta.$$

Else: Pick a root λ of the function $l^{(k)}(\lambda')$ in the domain $\mathbb{R}\backslash(-\xi^{(k)}, -\chi^{(k)})$. In the following two cases we consider the function f_λ on $[\chi^{(k)}, \xi^{(k)}]$.

* If $\lambda \neq -\chi^{(k)}$ then: Let $\eta = -f_\lambda(\chi^{(k)}) + 1$ where any positive constant could be chosen instead of 1. Define

$$X_h'^{(k)} := f_\lambda(X_h^{(k)}) + \eta, \quad X_g'^{(k)} := f_\lambda(X_g^{(k)}) + \eta.$$

* If $\lambda = -\chi^{(k)}$ then: Update $s^{(k)} = -1$. Let $\eta = -f_\lambda(\xi^{(k)}) - 1$ where any positive constant could be chosen instead of 1. Define

$$X_h'^{(k)} := X_g''^{(k)}, \quad X_g'^{(k)} := X_h''^{(k)},$$

where

$$X_h''^{(k)} := -f_\lambda(X_h^{(k)}) - \eta, \quad X_g''^{(k)} := -f_\lambda(X_g^{(k)}) - \eta$$

and make the replacement

$$\left|v^{(k)}\right\rangle \to \left|w^{(k)}\right\rangle$$
$$\left|w^{(k)}\right\rangle \to \left|v^{(k)}\right\rangle.$$

If we have $\lambda = -\chi^{(k)}$ or $-\xi^{(k)}$ it means that at least one of the matrices (among $X_g^{(k)}$ and $X_h^{(k)}$ under f_λ) would diverge. We must remove eigenvalues common to both matrices as isolating the divergence makes it easier to handle.

– **Remove spectral collision: If** $\lambda = -\chi^{(k)}$ or $\lambda = -\xi^{(k)}$ **then**

If it so happens that the coordinate and the probability associated is the same we must leave the associated vector unchanged (up to a relabelling). The following simply formalises this procedure and encodes the remaining non-trivial part into a problem of one less dimension.

1. **Idle point: If** for some j', j, we have $q^{(k)}_{g_{j'}} = q^{(k)}_{h_j}$ and $y^{(k)}_{g_{j'}} = y^{(k)}_{h_j}$ **then** the solution is given by

$$\left\{\left|u^{(k)}_h\right\rangle, \left|t^{(k)}_{h_1}\right\rangle, \left|t^{(k)}_{h_2}\right\rangle, \dots \left|t^{(k)}_{h_{k-1}}\right\rangle\right\} \overset{\text{componentwise}}{:=}$$
$$\left\{\left|t^{(k+1)}_{h_j}\right\rangle, \left|t^{(k+1)}_{h_1}\right\rangle, \left|t^{(k+1)}_{h_2}\right\rangle, \dots \left|t^{(k+1)}_{h_{j-1}}\right\rangle, \left|t^{(k+1)}_{h_{j+1}}\right\rangle, \dots \left|t^{(k+1)}_{h_k}\right\rangle\right\},$$

$$\bar{O}^{(k)} := \sum_{i=1}^{k} |a_i\rangle \left\langle t^{(k+1)}_{h_i}\right|,$$

where

$$\{|a_1\rangle, |a_2\rangle \dots |a_k\rangle\} \overset{\text{componentwise}}{:=}$$

$$\begin{cases} \left\{\left|t^{(k+1)}_{h_1}\right\rangle, \left|t^{(k+1)}_{h_2}\right\rangle, \dots \left|t^{(k+1)}_{h_{j-1}}\right\rangle, \left|t^{(k+1)}_{h_{j'}}\right\rangle, \left|t^{(k+1)}_{h_j}\right\rangle, \left|t^{(k+1)}_{h_{j+1}}\right\rangle, \right. & \\ \left. \dots, \left|t^{(k+1)}_{h_{j'-1}}\right\rangle, \left|t^{(k+1)}_{h_{j'+1}}\right\rangle \dots \left|t^{(k+1)}_{h_k}\right\rangle\right\} & j < j' \\ \left\{\left|t^{(k+1)}_{h_1}\right\rangle, \left|t^{(k+1)}_{h_2}\right\rangle, \dots \left|t^{(k+1)}_{h_{j'-1}}\right\rangle, \left|t^{(k+1)}_{h_{j'+1}}\right\rangle \dots \right. & \\ \left.\left|t^{(k+1)}_{h_{j-1}}\right\rangle, \left|t^{(k+1)}_{h_{j'}}\right\rangle, \left|t^{(k+1)}_{h_j}\right\rangle, \left|t^{(k+1)}_{h_{j+1}}\right\rangle \dots \left|t^{(k+1)}_{h_k}\right\rangle\right\} & j > j' \\ \left\{\left|t^{(k+1)}_{h_1}\right\rangle, \left|t^{(k+1)}_{h_2}\right\rangle, \dots \left|t^{(k+1)}_{h_k}\right\rangle\right\} & j = j', \end{cases}$$

and

$$X^{(k-1)}_h := \sum_{i \neq j} y^{(k)}_{h_i} \left|t^{(k+1)}_{h_i}\right\rangle \left\langle t^{(k+1)}_{h_i}\right|,$$

$$X^{(k-1)}_g := \bar{O}^{(k)} X^{(k)}_g \bar{O}^{(k)T} - y_{h_j} \left|t^{(k+1)}_{h_j}\right\rangle \left\langle t^{(k+1)}_{h_j}\right|,$$

$$\left|w^{(k-1)}\right\rangle = \mathcal{N}\left[\left|w^{(k)}\right\rangle - \sqrt{p_{h_j}}\left|t^{(k+1)}_{h_j}\right\rangle\right], \left|v^{(k-1)}\right\rangle = \mathcal{N}\left[\bar{O}^{(k)}\left|v^{(k)}\right\rangle - \sqrt{p_{h_j}}\left|t^{(k+1)}_{h_j}\right\rangle\right].$$

(This specifies $\underline{X}^{(k-1)} := \{X^{(k-1)}_h, X^{(k-1)}_g, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle\}$.)
Jump to **End**.

In this proof by x_{h_i} we mean y_{h_i}, similarly by x_{g_i} we mean y_{h_i}; we apologise for the inconvenience. We want to find an $O^{(k)}$ such that $X^{(k)}_h \geq O^{(k)} X^{(k)}_g O^{(k)T}$ and $O^{(k)}\left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle$. We do this in two stages. First, we re-arrange the entries of $X^{(k)}_g$ as $X^{\prime(k)}_g := O^{(k)}_p X^{(k)}_g O^{(k)T}_p$ and define $\left|v^{(k)}_p\right\rangle := O^{(k)}_p |v\rangle$ for an $O^{(k)}_p$ to be specified later. The re-arrangement will be such that $x_{g_{j'}}$ sits at the j, j location while the rest of the elements of $X^{\prime(k)}_g$ are arranged in the increasing order. Second, we solve our initial problem under the assumption that $j = j'$. The non-trivial part here would

be showing that we can take $O^{(k)}$ to have the form $\left(|j\rangle\langle j| + O^{(k-1)}\right)\bar{O}^{(k)}$ without loss of generality.
Let us start with the first step. We denote the orthogonal matrix $O = \sum_i |b_i\rangle\langle a_i|$ by $\{|a_1\rangle, |a_2\rangle, \ldots |a_k\rangle\} \rightarrow \{|b_1\rangle, |b_2\rangle, \ldots |b_k\rangle\}$ where $\{|b_i\rangle\}$ and $\{|a_i\rangle\}$ each constitute an orthonormal basis. Using this notation then for the case $j < j'$, we define $O_p^{(k)}$ by

$$\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \rightarrow$$
$$\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j'}}^{(k+1)}\right\rangle, \left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \ldots,\right.$$
$$\left.\left|t_{h_{j'-1}}^{(k+1)}\right\rangle, \left|t_{h_{j'+1}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\},$$

for $j' < j$ we define it by

$$\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \rightarrow$$
$$\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j'-1}}^{(k+1)}\right\rangle, \left|t_{h_{j'+1}}^{(k+1)}\right\rangle \ldots\right.$$
$$\left.\left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j'}}^{(k+1)}\right\rangle, \left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}$$

and if $j' = j$ we set $O_p^{(k)} = \mathbb{I}^{(k)}$.
For the second step, we solve the main problem under the assumption that $j' = j$. We are given $X_g^{'(k)} = \text{diag}\{x'_{g_1}, x'_{g_2} \ldots x'_{g_k}\}$ and $X_h^{(k)} = \text{diag}\{x_{h_1}, x_{h_2} \ldots x_{h_k}\}$ which are such that $x_{h_j} = x'_{g_j}$; $\left|v'^{(k)}\right\rangle \doteq (\sqrt{q'_{g_1}}, \sqrt{q'_{g_2}}, \ldots \sqrt{q'_{g_k}})^T$ and $\left|w^{(k)}\right\rangle \doteq (\sqrt{q_{h_1}}, \sqrt{q_{h_2}}, \ldots \sqrt{q_{h_k}})^T$ are such that $q_{h_j} = q'_{g_j}$. Let us define the matrix instance to be $\underline{X}'^{(k)} = \{X_h^{(k)}, X_g'^{(k)}, \left|v'^{(k)}\right\rangle, \left|w^{(k)}\right\rangle\}$. We have to find an $O'^{(k)}$ such that $X_h^{(k)} \geq O'^{(k)} X_g'^{(k)} O'^{(k)T}$ and $O'^{(k)}\left|v'^{(k)}\right\rangle = |w\rangle$. Let $\underline{X}'^{(k-1)} = \left\{X_h^{(k-1)}, X_g'^{(k-1)}, \left|v'^{(k-1)}\right\rangle, \left|w^{(k-1)}\right\rangle\right\}$ be the matrix instance obtained after removing the j^{th} entry from the vectors, viz. $\left|v'^{(k-1)}\right\rangle := \sum_{i\neq j}\sqrt{q'_{g_i}}\left|t_{h_i}^{(k+1)}\right\rangle$, $\left|w^{(k-1)}\right\rangle := \sum_{i\neq j}\sqrt{q_{h_i}}\left|t_{h_i}^{(k+1)}\right\rangle$ and similarly defining $X_g'^{(k-1)} = \text{diag}\{x'_{g_1}, x'_{g_2} \ldots x'_{g_{j-1}}, x'_{g_{j+1}}, \ldots x_{g_k}\}$, $X_h^{(k-1)} = \text{diag}\left\{x_{h_1}, x_{h_2} \ldots x_{h_{j-1}}, x_{h_{j+1}}, \ldots x_{h_k}\right\}$. Note that $a^{(k)} = a^{(k-1)}$ as the j^{th} point gets cancelled. This means that if there is an $O'^{(k)}$ satisfying the aforementioned constraints $a^{(k)}$ is EBRM on the spectral domain of $\underline{X}^{(k)}$. Since $a^{(k)} = a^{(k-1)}$ we know that $a^{(k-1)}$ is also EBRM on the same domain. From Lemma 130 (we will justify that k is large enough separately) we conclude that there must also exist an $O'^{(k-1)}$ which satisfies $X_h^{(k-1)} \geq O'^{(k-1)} X_g'^{(k-1)} O'^{(k-1)T}$ and $O'^{(k-1)}\left|v'^{(k-1)}\right\rangle = \left|w^{(k)}\right\rangle$.
With all this in place we can claim that without loss of generality we can write $O'^{(k)} = \left|t_{h_j}\right\rangle\left\langle t_{h_j}\right| + O'^{(k-1)}$ because if we can find some other $\tilde{O}'^{(k)}$ which satisfies the required constraints then there exists an $O'^{(k-1)}$ which satisfies the corresponding constraints in the smaller dimension and that means we can show $O'^{(k)}$ also satisfies the required constraints,

$$X_h^{(k)} = x_{h_j}\left|t_{h_j}^{(k+1)}\right\rangle\left\langle t_{h_j}^{(k+1)}\right| + X_h^{(k-1)} \geq$$
$$x_{g_j}\left|t_{h_j}^{(k+1)}\right\rangle\left\langle t_{h_j}^{(k+1)}\right| + O'^{(k-1)} X_g'^{(k-1)} O'^{(k-1)} =$$
$$\left(\left|t_{h_j}^{(k+1)}\right\rangle\left\langle t_{h_j}^{(k+1)}\right| + O'^{(k-1)}\right) X_g'^{(k)} \left(\left|t_{h_j}^{(k+1)}\right\rangle\left\langle t_{h_j}^{(k+1)}\right| + O'^{(k-1)}\right)^T =$$
$$O'^{(k)} X_g'^{(k)} O'^{(k)T},$$

along with

$$O'^{(k)}\left|v'^{(k)}\right\rangle = \sqrt{q'_{g_j}}\left|t_{h_j}^{(k+1)}\right\rangle + O'^{(k-1)}\left|v'^{(k-1)}\right\rangle = \sqrt{q'_{g_j}}\left|t_{h_j}^{(k+1)}\right\rangle + \left|w^{(k-1)}\right\rangle = \left|w^{(k-1)}\right\rangle.$$

It remains to combine the two steps to produce the matrix $\bar{O}^{(k)}$, the vectors $\left\{\left|n_h^{(k)}\right\rangle, \left\{\left|t_{h_i}^{(k)}\right\rangle\right\}\right\}$, along with $\underline{\mathrm{X}}^{(k-1)}$. We use $X_g'^{(k)} = O_p^{(k)} X_g^{(k)} O_p^{(k)T}$ from the first step and substitute it in the inequality which we showed would hold, i.e.

$$X_h^{(k)} \geq O'^{(k)} X_g'^{(k)} O'^{(k)T} = O'^{(k)} O_p^{(k)} X_g O_p^{(k)T} O'^{(k)T}$$

and using $O_p^{(k)}\left|v^{(k)}\right\rangle = \left|v'^{(k)}\right\rangle$ we have

$$O'^{(k)}\left|v'^{(k)}\right\rangle = O'^{(k)} O_p^{(k)}\left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle.$$

Comparing the inequality to the form $X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T}$, $O^{(k)}\left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle$ for

$$O^{(k)} = \left(\left|n_h^{(k)}\right\rangle\left\langle n_h^{(k)}\right| + O^{(k-1)}\right)\bar{O}^{(k)}$$

we get $\bar{O}^{(k)} = O_p^{(k)}$, $\left|n_h^{(k)}\right\rangle = \left|t_{h_j}^{(k+1)}\right\rangle$ and $O^{(k-1)} = O'^{(k-1)}$. Note that this $O^{(k)}$ is consistent with comparing the equality with $O^{(k)}\left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle$. The basis for the sub-problem, i.e. the $(k-1)$ dimensional problem, was the same as before except for the fact that we removed $\left|t_{h_j}^{(k+1)}\right\rangle$. Thus we define $\left\{\left|t_{h_1}^{(k)}\right\rangle, \left|t_{h_2}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} = \left\{t_{h_1}^{(k+1)}, t_{h_2}^{(k+1)} \ldots t_{h_{j-1}}^{(k+1)}, t_{h_{j+1}}^{(k+1)}, \ldots t_{h_k}^{(k+1)}\right\}$. Identifying

$$\underline{\mathrm{X}}^{(k-1)} = \left\{X_h^{(k-1)}, X_g^{(k-1)}, \left|v^{(k-1)}\right\rangle, \left|w^{(k-1)}\right\rangle\right\}$$

with

$$\underline{\mathrm{X}}'^{(k-1)} = \left\{X_h^{(k-1)}, X_g'^{(k-1)}, \left|v'^{(k-1)}\right\rangle, \left|w^{(k-1)}\right\rangle\right\}$$

completes the argument since $O^{(k-1)}$ was already identified with $O'^{(k-1)}$ so we are just labelling here.

2. **Final Extra**: **If** for some j, j' we have $q_{g_{j'}}^{(k)} > q_{h_j}^{(k)}$ and $y_{g_{j'}}^{(k)} = y_{h_j}^{(k)}$ **then** the solution is given by $\underline{\mathrm{X}}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)}\left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|$, $X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)}\left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|$, $\left|v^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1}\sqrt{q_{g_i}^{(k-1)}}\left|t_{h_i}^{(k)}\right\rangle\right]$, $\left|w^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1}\sqrt{q_{h_i}^{(k-1)}}\left|t_{h_i}^{(k)}\right\rangle\right]$ where the co-ordinates and weights are given by

$$\begin{aligned}
\left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)}, q_{h_2}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_{j+1}}^{(k)}, \ldots q_{h_k}^{(k)}\right\} \\
\left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}, q_{g_{j'+1}}^{(k)}, q_{g_{j'+2}}^{(k)} \ldots q_{g_k}^{(k)}\right\} \\
\left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{g_2}^{(k)}, \ldots y_{g_k}^{(k)}\right\} \\
\left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{j-1}}^{(k)}, y_{h_{j+1}}^{(k)} \ldots, y_{h_k}^{(k)}\right\},
\end{aligned}$$

the basis is given by

$$\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=}$$
$$\left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}.$$

The orthogonal matrices are given by $\bar{O}_h^{(k)} := \sum \left|t_{h_i}^{(k+1)}\right\rangle \langle a_i|$ where

$$\{|a_1\rangle, \ldots |a_k\rangle\} \to \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\},$$

$\bar{O}_g^{(k)} := \tilde{O}^{(k)} \bar{O}_h^{(k)}$ where

$$\begin{aligned}\tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{h_j}^{(k)}}\left|u_h^{(k)}\right\rangle + \sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}}\left|t_{h_{j'}}^{(k)}\right\rangle\right]\mathcal{N}\left[\sqrt{q_{g_1}^{(k)}}\left\langle u_h^{(k)}\right| + \sqrt{q_{g_{j'}}^{(k)}}\left\langle t_{h_{j'}}^{(k)}\right|\right] \\ &+ \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}}\left|u_h^{(k)}\right\rangle - \sqrt{q_{h_j}^{(k)}}\left|t_{h_{j'}}^{(k)}\right\rangle\right]\mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}}\left\langle u_h^{(k)}\right| - \sqrt{q_{g_1}^{(k)}}\left\langle t_{h_{j'}}^{(k)}\right|\right] \\ &+ \sum_{i\in\{1,\ldots k\}\setminus j'}\left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|.\end{aligned}$$

Jump to **End**.

We are given $\underline{\mathrm{X}}^{(k)} = (X_h^{(k)}, X_g^{(k)}, |w^{(k)}\rangle, |v^{(k)}\rangle)$ where $X_h^{(k)} = \sum_{i=1}^k y_{h_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle\left\langle t_{h_i}^{(k+1)}\right|$, $X_g^{(k)} = \sum_{i=1}^k y_{g_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle\left\langle t_{h_i}^{(k+1)}\right|$, $|v^{(k)}\rangle = \sum_{i=1}^k q_{g_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle$, $|w^{(k)}\rangle = \sum_{i=1}^k q_{h_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle$ which means the corresponding function instance $\underline{\mathrm{x}}^{(k)} = (h^{(k)}, g^{(k)}, a^{(k)})$ where, in particular we have, $a^{(k)} = \sum_{i\in\{1,\ldots k\}\setminus j} q_{h_i}^{(k)}[y_{h_i}] - \sum_{i\in\{1,\ldots k\}\setminus j'} q_{g_i}^{(k)}[y_{g_i}] - (q_{g_{j'}}^{(k)} - q_{h_j}^{(k)})[y_{h_j}]$. Since we assume $\underline{\mathrm{X}}^{(k)}$ has a solution it follows that $a^{(k)}$ is $[\chi, \xi]$ valid. Thus the transition $g^{(k-1)} := a_-^{(k)} \to a_+^{(k)} =: h^{(k-1)}$ is also $[\chi, \xi]$ valid where $g^{(k-1)}$ comprises $n_g^{(k-1)} = n_g^{(k)}$ points and $h^{(k-1)}$ comprises $n_h^{(k-1)} = n_h^{(k)} - 1$ points (using the attributes corresponding to the function instance $(h^{(k-1)}, g^{(k-1)}, h^{(k-1)} - g^{(k-1)})$; The notation would be of the form $g = \sum_{i=1}^{n_g} p_{g_i}[x_{g_i}]$ and $h = \sum_{i=1}^{n_h} p_{h_i}[x_{h_i}]$). Since $k = n_g^{(k)} + n_h^{(k)} - 1$ the aforesaid relation yields $k - 1 = n_g^{(k-1)} + n_h^{(k-1)} - 1$. We conclude that $\underline{\mathrm{X}}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, |w^{(k-1)}\rangle, |v^{(k-1)}\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|$, $X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle\left\langle t_{h_i}^{(k)}\right|$, $|v^{(k-1)}\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1}\sqrt{q_{g_i}^{(k-1)}}\left|t_{h_i}^{(k)}\right\rangle\right]$, $|w^{(k-1)}\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1}\sqrt{q_{h_i}^{(k-1)}}\left|t_{h_i}^{(k)}\right\rangle\right]$ has a solution for

$$\begin{aligned}
&\left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} \overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)}, q_{h_2}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_{j+1}}^{(k)}, \ldots q_{h_k}^{(k)}\right\} \\
&\left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} \overset{\text{componentwise}}{=} \left\{q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}, q_{g_{j'+1}}^{(k)}, q_{g_{j'+2}}^{(k)} \ldots q_{g_k}^{(k)}\right\} \\
&\left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} \overset{\text{componentwise}}{=} \left\{y_{g_2}^{(k)}, \ldots y_{g_k}^{(k)}\right\} \\
&\left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} \overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{j-1}}^{(k)}, y_{h_{j+1}}^{(k)} \ldots, y_{h_k}^{(k)}\right\}
\end{aligned}$$

as the corresponding function instance $\underline{\mathrm{x}}^{(k-1)}$ is indeed given by $(h^{(k-1)}, g^{(k-1)}, a^{(k-1)} = a^{(k)})$. Here $\{\left|t_{h_i}^{(k)}\right\rangle\}$ constitute an orthonormal basis

which we relate to $\left|t_{h_i}^{(k+1)}\right\rangle$ shortly. We used the fact that $q_{g_1}^{(k)} = 0$ as $y_{g_1}^{(k)} = \chi$ (To see this note that $k-1 > n_g^{(k-1)}$ which means that many q_{g_i} are zero; by convention we write the smallest eigenvalue, χ first to increase the matrix size so the first $i = 1, 2 \ldots \left(k-1-n_g^{(k-1)}\right)$ q_is are zero.). This means that there must exist an $O^{(k-1)}$ which solves $\underline{\mathrm{X}}^{(k-1)}$.

Let us take a moment to note the following basis change manoeuvre. Note that $X_h' \geq O' X_g' O'^T$ with $O' \left|v'\right\rangle = \left|w'\right\rangle$ is equivalent to $X_h \geq O X_g O^T$ with $O \left|v\right\rangle = \left|w\right\rangle$ where $O = \bar{O}_h^T O' \bar{O}_g$, $\bar{O}_g \left|v\right\rangle = \left|v'\right\rangle$, $\bar{O}_h \left|w\right\rangle = \left|w'\right\rangle$, $\bar{O}_h X_h \bar{O}_h^T = X_h'$, $\bar{O} X_g \bar{O}_g^T = X_g'$ which is easy to see by a simple substitution.

We first expand the matrix $\underline{\mathrm{X}}^{(k-1)}$ to k dimensions as follows. We already had $X_h^{(k-1)} \geq O^{(k-1)} X_g^{(k-1)} O^{(k-1)T}$ with $O^{(k-1)} \left|v^{(k-1)}\right\rangle = \left|w^{(k-1)}\right\rangle$ which we expand as

$$\underbrace{y_{h_j}^{(k)} \left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + X_h^{(k-1)}}_{:=X_h'^{(k)}} \geq$$

$$\underbrace{\left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right)}_{:=O'^{(k)}} \underbrace{\left(y_{h_j}^{(k)} \left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + X_g^{(k-1)}\right)}_{:=X_g'^{(k)}} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right)^T$$

with $\left|v'^{(k)}\right\rangle = \mathcal{N}\left[\sqrt{q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle + \left|v^{(k-1)}\right\rangle\right]$ and $\left|w'^{(k)}\right\rangle = \mathcal{N}\left[\sqrt{q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle + \left|w^{(k-1)}\right\rangle\right]$. Note that the matrix instance $\underline{\mathrm{X}}'^{(k)} := (X_h'^{(k)}, X_g'^{(k)}, \left|v'^{(k)}\right\rangle, \left|w'^{(k)}\right\rangle)$ yields $\underline{\mathrm{x}}'^{(k)} = \underline{\mathrm{x}}^{(k)}$. We can now use the equivalence we pointed out above to establish a relation between $X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T}$ and $X_h'^{(k)} \geq O'^{(k)} X_g'^{(k)} O'^{(k)T}$ by finding $\bar{O}_g$ and $\bar{O}_h$. We define, somewhat arbitrarily,

$$\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=}$$
$$\left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}.$$

We require $\bar{O}_h^{(k)} \left|w^{(k)}\right\rangle$ to be $\left|w'^{(k)}\right\rangle$. This is simply a permutation matrix given by $\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \to \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\}$. Note that this yields $\bar{O}_h^{(k)T} X_h'^{(k)} \bar{O}_h^{(k)} = X_h^{(k)}$. It remains to find $\bar{O}_g^{(k)}$ which we demand must satisfy $\bar{O}_g^{(k)} \left|v^{(k)}\right\rangle = \left|v'^{(k)}\right\rangle$. Observe first that $\bar{O}_h^{(k)} \left|v^{(k)}\right\rangle = \sqrt{q_{g_1}^{(k)}} \left|u_h^{(k)}\right\rangle + \sum_{i=2}^k \sqrt{q_{g_i}^{(k)}} \left|t_{h_{i-1}}^{(k)}\right\rangle$. We must now apply

$$\begin{aligned}
\tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle + \sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}} \left|t_{h_{j'}}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_1}^{(k)}} \left\langle u_h^{(k)}\right| + \sqrt{q_{g_{j'}}^{(k)}} \left\langle t_{h_{j'}}^{(k)}\right|\right] \\
&+ \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)} - q_{h_j}^{(k)}} \left|u_h^{(k)}\right\rangle - \sqrt{q_{h_j}^{(k)}} \left|t_{h_{j'}}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left\langle u_h^{(k)}\right| - \sqrt{q_{g_1}^{(k)}} \left\langle t_{h_{j'}}^{(k)}\right|\right] \\
&+ \sum_{i \in \{1,\ldots k\} \setminus j'} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|
\end{aligned}$$

to get $\bar{O}_g^{(k)} \left|v^{(k)}\right\rangle = \left|v'^{(k)}\right\rangle$ where we defined $\bar{O}_g^{(k)} := \tilde{O}^{(k)} \bar{O}_h^{(k)}$. (Note the expression could be simplified by using $q_{g_1} = 0$ which in fact is necessary for probability

conservation.) Using $y_{h_j}^{(k)} = y_{g_{j'}}^{(k)}$ we can also see that $\bar{O}_g^{(k)T} X_g'^{(k)} \bar{O}_g^{(k)}$ is essentially $X_g^{(k)}$ with $\chi^{(k)}$ at $\left|t_{h_1}^{(k+1)}\right\rangle$ replaced by $y_{g_{j'}} (= y_{h_j})$. One can conclude therefore that $X_g'^{(k)} \geq \bar{O}_g^{(k)} X_g^{(k)} \bar{O}_g^{(k)T}$. Following the substitution manoeuvre we have

$$\begin{aligned} & X_h'^{(k)} \geq O'^{(k)} X_g'^{(k)} O'^{(k)T} \geq O'^{(k)} \bar{O}_g^{(k)} X_g^{(k)} \bar{O}_g^{(k)T} O'^{(k)T} \\ \iff & \bar{O}_h^{(k)T} X_h'^{(k)} \bar{O}_h^{(k)} \geq \underbrace{\bar{O}_h^{(k)T} O'^{(k)} \bar{O}_g^{(k)}}_{:=O^{(k)}} X_g^{(k)} \bar{O}_g^{(k)T} O'^{(k)T} \bar{O}_h^{(k)} \\ \iff & X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T} \end{aligned}$$

and similarly

$$\begin{aligned} & O'^{(k)} \left|v'^{(k)}\right\rangle = \left|w'^{(k)}\right\rangle \\ \iff & O'^{(k)} \bar{O}_g^{(k)} \left|v^{(k)}\right\rangle = \bar{O}_h^{(k)} \left|w^{(k)}\right\rangle \\ \iff & O^{(k)} \left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle. \end{aligned}$$

This completes the proof.

3. **Initial Extra**: **If** for some j, j' we have $q_{g_{j'}}^{(k)} < q_{h_j}^{(k)}$ and $y_{g_{j'}}^{(k)} = y_{h_j}^{(k)}$ **then** the solution is given by $\underline{X}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, \left|w^{(k-1)}\right\rangle, \left|v^{(k-1)}\right\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|$, $X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|$, $\left|v^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{g_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$, $\left|w^{(k-1)}\right\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{h_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$ where the co-ordinates and weights are given by

$$\begin{aligned} \left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}, q_{h_{j+1}}^{(k)}, q_{h_{j+2}}^{(k)} \ldots q_{h_{k-1}}^{(k)}\right\} \\ \left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{g_1}^{(k)}, q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'+1}}^{(k)}, \ldots q_{g_k}^{(k)}\right\} \\ \left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{g_1}^{(k)}, \ldots y_{g_{j'-1}}^{(k)}, y_{g_{j'+1}}^{(k)} \ldots, y_{g_k}^{(k)}\right\} \\ \left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{k-1}}^{(k)}\right\}, \end{aligned}$$

the basis is given by

$$\begin{aligned} & \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=} \\ & \left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}. \end{aligned}$$

The orthogonal matrices are given by $\bar{O}_h^{(k)} := \tilde{O}^{(k)} \sum |a_i\rangle \left\langle t_{h_i}^{(k+1)}\right|$ where

$$\{|a_1\rangle, \ldots |a_k\rangle\} \overset{\text{componentwise}}{=} \left\{\left|t_{h_1}^{(k)}\right\rangle, \left|t_{h_2}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle, \left|u_h^{(k)}\right\rangle\right\}.$$

$$\begin{aligned} \tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle + \sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}} \left|t_{h_j}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{h_k}^{(k)}} \left\langle u_h^{(k)}\right| + \sqrt{q_{g_j}^{(k)}} \left\langle t_{h_j}^{(k)}\right|\right] \\ & + \mathcal{N}\left[\sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle - \sqrt{q_{g_{j'}}^{(k)}} \left|t_{h_j}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_j}^{(k)}} \left\langle u_h^{(k)}\right| - \sqrt{q_{h_k}^{(k)}} \left\langle t_{h_j}^{(k)}\right|\right] \\ & + \sum_{i \in \{1, \ldots k\} \setminus j} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right| \end{aligned}$$

and $\bar{O}_h^{(k)}$ is given by the basis change $\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \rightarrow \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\}$.
Jump to **End**.

This proof will be very similar to the previous one. We are given $\underline{\mathrm{X}}^{(k)} = (X_h^{(k)}, X_g^{(k)}, |w^{(k)}\rangle, |v^{(k)}\rangle)$ where $X_h^{(k)} = \sum_{i=1}^k y_{h_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle \left\langle t_{h_i}^{(k+1)}\right|$, $X_g^{(k)} = \sum_{i=1}^k y_{g_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle \left\langle t_{h_i}^{(k+1)}\right|$, $|v^{(k)}\rangle = \sum_{i=1}^k q_{g_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle$, $|w^{(k)}\rangle = \sum_{i=1}^k q_{h_i}^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle$ which means the corresponding function instance $\underline{\mathrm{x}}^{(k)} = (h^{(k)}, g^{(k)}, a^{(k)})$ where, in particular we have,

$$a^{(k)} = \sum_{i \in \{1,\ldots k\} \setminus j} q_{h_i}^{(k)}[y_{h_i}] + (q_{h_j}^{(k)} - q_{g_{j'}}^{(k)})[y_{h_j}] - \sum_{i \in \{1,\ldots k\} \setminus j'} q_{g_i}^{(k)}[y_{g_i}].$$

Since we assume $\underline{\mathrm{X}}^{(k)}$ has a solution it follows that $a^{(k)}$ is $[\chi, \xi]$ valid. Thus the transition $g^{(k-1)} := a_-^{(k)} \rightarrow a_+^{(k)} =: h^{(k-1)}$ is also $[\chi, \xi]$ valid where $g^{(k-1)}$ comprises $n_g^{(k-1)} = n_g^{(k)} - 1$ points and $h^{(k-1)}$ comprises $n_h^{(k-1)} = n_h^{(k)}$ points (using the attributes corresponding to the function instance $(h^{(k-1)}, g^{(k-1)}, h^{(k-1)} - g^{(k-1)})$; The notation would be of the form $g = \sum_{i=1}^{n_g} p_{g_i}[x_{g_i}]$ and $h = \sum_{i=1}^{n_h} p_{h_i}[x_{h_i}]$). Since $k = n_g^{(k)} + n_h^{(k)} - 1$ the aforesaid relation yields $n_g^{(k-1)} + n_h^{(k-1)} - 1 = k - 1$. We conclude that $\underline{\mathrm{X}}^{(k-1)} := (X_h^{(k-1)}, X_g^{(k-1)}, |w^{(k-1)}\rangle, |v^{(k-1)}\rangle)$ where $X_h^{(k-1)} = \sum_{i=1}^{k-1} y_{h_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|$, $X_g^{(k-1)} = \sum_{i=1}^{k-1} y_{g_i}^{(k-1)} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|$, $|v^{(k-1)}\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{g_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$, $|w^{(k-1)}\rangle = \mathcal{N}\left[\sum_{i=1}^{k-1} \sqrt{q_{h_i}^{(k-1)}} \left|t_{h_i}^{(k)}\right\rangle\right]$ have a solution for

$$\begin{aligned}
\left\{q_{h_1}^{(k-1)}, \ldots q_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{h_1}^{(k)} \ldots, q_{h_{j-1}}^{(k)}, q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}, q_{h_{j+1}}^{(k)}, q_{h_{j+2}}^{(k)} \ldots q_{h_{k-1}}^{(k)}\right\} \\
\left\{q_{g_1}^{(k-1)}, \ldots q_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{q_{g_1}^{(k)}, q_{g_2}^{(k)} \ldots, q_{g_{j'-1}}^{(k)}, q_{g_{j'+1}}^{(k)}, \ldots q_{g_k}^{(k)}\right\} \\
\left\{y_{g_1}^{(k-1)}, \ldots y_{g_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{g_1}^{(k)}, \ldots y_{g_{j'-1}}^{(k)}, y_{g_{j'+1}}^{(k)} \ldots, y_{g_k}^{(k)}\right\} \\
\left\{y_{h_1}^{(k-1)}, \ldots y_{h_{k-1}}^{(k-1)}\right\} &\overset{\text{componentwise}}{=} \left\{y_{h_1}^{(k)}, \ldots y_{h_{k-1}}^{(k)}\right\},
\end{aligned}$$

as the corresponding function instance $\underline{\mathrm{x}}^{(k-1)}$ is indeed given by $(h^{(k-1)}, g^{(k-1)}, a^{(k-1)} = a^{(k)})$. Here $\{\left|t_{h_i}^{(k)}\right\rangle\}$ constitute an orthonormal basis which we relate to $\left|t_{h_i}^{(k+1)}\right\rangle$ shortly. We used the fact that $q_{h_k}^{(k)} = 0$ as $y_{h_k}^{(k)} = \xi$. (To see this note that $k-1 > n_h^{(k-1)}$ which means that many q_{h_i} are zero; by convention we write the smallest eigenvalue, x_{h_1} first all the way till $x_{h_{n_h}}$ and then to increase the matrix size we append zeros so the $i = n_h, n_h + 1 \ldots k$ yield $q_{h_i} = 0$.) This means that there must exist an $O^{(k-1)}$ which solves $\underline{\mathrm{X}}^{(k-1)}$.
Let us take a moment to note the following basis change manoeuvre. $X_h' \geq O' X_g' O'^T$ with $O'|v'\rangle = |w'\rangle$ is equivalent to $X_h \geq O X_g O^T$ with $O|v\rangle = |w\rangle$ where $O = \bar{O}_h^T O' \bar{O}_g$, $\bar{O}_g|v\rangle = |v'\rangle$, $\bar{O}_h|w\rangle = |w'\rangle$, $\bar{O}_h X_h \bar{O}_h^T = X_h'$, $\bar{O} X_g \bar{O}_g^T = X_g'$ which is easy to see by a simple substitution.
We first expand the matrix $\underline{\mathrm{X}}^{(k-1)}$ to k dimensions as follows. We already had $X_h^{(k-1)} \geq$

$O^{(k-1)} X_g^{(k-1)} O^{(k-1)T}$ with $O^{(k-1)} \left|v^{(k-1)}\right\rangle = \left|w^{(k-1)}\right\rangle$ which we expand as

$$\underbrace{y_{h_j}^{(k)} \left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + X_h^{(k-1)}}_{:=X_h^{\prime(k)}} \geq$$

$$\underbrace{\left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right)}_{:=O^{\prime(k)}} \underbrace{\left(y_{h_j}^{(k)} \left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + X_g^{(k-1)}\right)}_{:=X_g^{\prime(k)}} \left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right)^T$$

with $\left|v^{\prime(k)}\right\rangle = \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle + \left|v^{(k-1)}\right\rangle\right]$ and $\left|w^{\prime(k)}\right\rangle = \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle + \left|w^{(k-1)}\right\rangle\right]$. Note that the matrix instance $\underline{X}^{\prime(k)} := (X_h^{\prime(k)}, X_g^{\prime(k)}, \left|v^{\prime(k)}\right\rangle, \left|w^{\prime(k)}\right\rangle)$ yields $\underline{x}^{\prime(k)} = \underline{x}^{(k)}$. We can now use the equivalence we pointed out above to establish a relation between $X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T}$ and $X_h^{\prime(k)} \geq O^{\prime(k)} X_g^{\prime(k)} O^{\prime(k)T}$ by finding $\bar{O}_g$ and $\bar{O}_h$. We define, somewhat arbitrarily,

$$\left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\} \overset{\text{componentwise}}{=}$$
$$\left\{\left|t_{h_j}^{(k+1)}\right\rangle, \left|t_{h_1}^{(k+1)}\right\rangle, \left|t_{h_2}^{(k+1)}\right\rangle, \ldots \left|t_{h_{j-1}}^{(k+1)}\right\rangle, \left|t_{h_{j+1}}^{(k+1)}\right\rangle, \left|t_{h_{j+2}}^{(k+1)}\right\rangle \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\}.$$

We require $\bar{O}_g^{(k)} \left|v^{(k)}\right\rangle$ to be $\left|v^{\prime(k)}\right\rangle$. This is simply a permutation matrix given by $\left\{\left|t_{h_1}^{(k+1)}\right\rangle, \ldots \left|t_{h_k}^{(k+1)}\right\rangle\right\} \rightarrow \left\{\left|u_h^{(k)}\right\rangle, \left|t_{h_1}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle\right\}$. Note that this yields $\bar{O}_g^{(k)T} X_g^{\prime(k)} \bar{O}_g^{(k)} = X_g^{(k)}$ as $y_{h_j}^{(k)} = y_{g_{j'}}^{(k)}$. It remains to find $\bar{O}_h^{(k)}$ which we demand must satisfy $\bar{O}_h^{(k)} \left|w^{(k)}\right\rangle = \left|w^{\prime(k)}\right\rangle$. Let us define $\bar{O}_h^{(k)} = \tilde{O}^{(k)} \left(\sum_{i=1}^k |a_i\rangle \left\langle t_{h_i}^{(k+1)}\right|\right)$. Observe that for $\tilde{O}^{(k)} = \mathbb{I}$ we have $\bar{O}_h^{(k)} \left|w^{(k)}\right\rangle = q_{h_k}^{(k)} \left|u_h^{(k)}\right\rangle + \sum_{i=1}^{k-1} q_{h_i}^{(k)} \left|t_{h_i}^{(k)}\right\rangle$ where

$$\{|a_1\rangle, \ldots |a_k\rangle\} \overset{\text{componentwise}}{=} \left\{\left|t_{h_1}^{(k)}\right\rangle, \left|t_{h_2}^{(k)}\right\rangle \ldots \left|t_{h_{k-1}}^{(k)}\right\rangle, \left|u_h^{(k)}\right\rangle\right\}.$$

If we define

$$\begin{aligned}\tilde{O}^{(k)} :=& \mathcal{N}\left[\sqrt{q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle + \sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}} \left|t_{h_j}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{h_k}^{(k)}} \left\langle u_h^{(k)}\right| + \sqrt{q_{g_j}^{(k)}} \left\langle t_{h_j}^{(k)}\right|\right] \\ &+ \mathcal{N}\left[\sqrt{q_{h_j}^{(k)} - q_{g_{j'}}^{(k)}} \left|u_h^{(k)}\right\rangle - \sqrt{q_{g_{j'}}^{(k)}} \left|t_{h_j}^{(k)}\right\rangle\right] \mathcal{N}\left[\sqrt{q_{g_j}^{(k)}} \left\langle u_h^{(k)}\right| - \sqrt{q_{h_k}^{(k)}} \left\langle t_{h_j}^{(k)}\right|\right] \\ &+ \sum_{i \in \{1,\ldots k\} \setminus j} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{h_i}^{(k)}\right|\end{aligned}$$

we get $\bar{O}_h^{(k)} \left|w^{(k)}\right\rangle = \left|w^{\prime(k)}\right\rangle$ as desired. We can also see that $\bar{O}_h^{(k)T} X_g^{\prime(k)} \bar{O}_h^{(k)}$ is essentially X_g with $\xi^{(k)}$ at $\left|t_{h_k}^{(k+1)}\right\rangle$ replaced by y_{h_j}. We therefore conclude that $X_g^{\prime(k)} \geq \bar{O}_g^{(k)} X_g^{(k)} \bar{O}_g^{\prime(k)}$. Following the substitution manoeuvre we have

$$\begin{aligned} & X_h^{\prime(k)} \geq O^{\prime(k)} X_g^{\prime(k)} O^{\prime(k)T} \geq O^{\prime(k)} \bar{O}_g^{(k)} X_g^{(k)} \bar{O}_g^{(k)T} O^{\prime(k)T} \\ \iff & \bar{O}_h^{(k)T} X_h^{\prime(k)} \bar{O}_h^{(k)} \geq \underbrace{\bar{O}_h^{(k)T} O^{\prime(k)} \bar{O}_g^{(k)}}_{:=O^{(k)}} X_g^{(k)} \bar{O}_g^{(k)T} O^{\prime(k)T} \bar{O}_h^{(k)} \\ \iff & X_h^{(k)} \geq O^{(k)} X_g^{(k)} O^{(k)T} \end{aligned}$$

and similarly

$$\begin{aligned} & O'^{(k)} \left|v'^{(k)}\right\rangle = \left|w'^{(k)}\right\rangle \\ \iff & O'^{(k)} \bar{O}_g^{(k)} \left|v^{(k)}\right\rangle = \bar{O}_h^{(k)} \left|w^{(k)}\right\rangle \\ \iff & O^{(k)} \left|v^{(k)}\right\rangle = \left|w^{(k)}\right\rangle . \end{aligned}$$

This completes the proof.

– **Evaluate the Reverse Weingarten Map**:

1. Consider the point $\left|w^{(k)}\right\rangle / \sqrt{\left\langle w^{(k)}\right| X_h'^{(k)} \left|w^{(k)}\right\rangle}$ on the ellipsoid $X_h'^{(k)}$. Evaluate the normal at this point as $\left|u_h^{(k)}\right\rangle = \mathcal{N}\left(\sum_{i=1}^{n_h^{(k)}} \sqrt{p_{h_i}^{(k)}} x_{h_i}'^{(k)} \left|t_{h_i}^{(k+1)}\right\rangle\right)$. Similarly evaluate $\left|u_g^{(k)}\right\rangle$, the normal at the point $\left|v^{(k)}\right\rangle / \sqrt{\left\langle w^{(k)}\right| X_g'^{(k)} \left|w^{(k)}\right\rangle}$ on the ellipsoid $X_g'^{(k)}$.
2. Recall that for a given diagonal matrix $X = \sum_i y_i \left|i\right\rangle \left\langle i\right| > 0$ and normal vector $\left|u\right\rangle = \sum_i u_i \left|i\right\rangle$ the Reverse Weingarten map is given by $W_{ij} = \left(-\frac{y_j^{-1} y_i^{-1} u_i u_j}{r^2} + y_i^{-1}\delta_{ij}\right)$ where $r = \sqrt{\sum y_i^{-1} u_i^2}$. Evaluate the Reverse Weingarten maps $W_h'^{(k)}$ and $W_g'^{(k)}$ along $\left|u_h^{(k)}\right\rangle$ and $\left|u_g^{(k)}\right\rangle$ respectively.
3. Find the eigenvectors and eigenvalues of the Reverse Weingarten maps. The eigenvectors of W_h' form the h tangent (and normal) vectors $\left\{\left\{\left|t_{h_i}^{(k)}\right\rangle\right\}, \left|u_h^{(k)}\right\rangle\right\}$. The corresponding radii of curvature are obtained from the eigenvalues $\left\{\{r_{h_i}^{(k)}\}, 0\right\} = \left\{\{c_{h_i}^{(k)-1}\}, 0\right\}$ which are inverses of the curvature values. The tangents are labelled in the decreasing order of radii of curvature (increasing order of curvature). Similarly for the g tangent (and normal) vectors. Fix the sign freedom in the eigenvectors by requiring $\left\langle t_{h_i}^{(k)} | w^{(k)}\right\rangle \geq 0$ and $\left\langle t_{g_i}^{(k)} | v^{(k)}\right\rangle \geq 0$.

– **Finite Method: If** $\lambda \neq -\xi^{(k)}$ and $\lambda \neq -\chi^{(k)}$, i.e. if it is the finite case **then**

1. $\bar{O}^{(k)} := \left|u_h^{(k)}\right\rangle \left\langle u_g^{(k)}\right| + \sum_{i=1}^{k-1} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{g_i}^{(k)}\right|$
2. $\left|v^{(k-1)}\right\rangle := \bar{O}^{(k)} \left|v^{(k)}\right\rangle - \left\langle u_h^{(k)}\right| \bar{O}^{(k)} \left|v^{(k)}\right\rangle \left|u_h^{(k)}\right\rangle$ and $\left|w^{(k-1)}\right\rangle := \left|w^{(k)}\right\rangle - \left\langle u_h^{(k)} | w^{(k)}\right\rangle \left|u_h^{(k)}\right\rangle$.
3. Define $X_h^{(k-1)} := \mathrm{diag}\{c_{h_1}^{(k)}, c_{h_2}^{(k)} \ldots, c_{h_{k-1}}^{(k)}\}$, $X_g^{(k-1)} := \mathrm{diag}\{c_{g_1}^{(k)}, c_{g_2}^{(k)} \ldots c_{g_{k-1}}^{(k)}\}$.
4. **Jump** to **End**.

Our first burden is to prove that $O^{(k)}$ must have the form $\left(\left|u_h^{(k)}\right\rangle \left\langle u_h^{(k)}\right| + O^{(k-1)}\right) \bar{O}^{(k)}$ for $\bar{O}^{(k)} := \left|u_h^{(k)}\right\rangle \left\langle u_g^{(k)}\right| + \sum_{i=1}^{k-1} \left|t_{h_i}^{(k)}\right\rangle \left\langle t_{g_i}^{(k)}\right|$ if $O^{(k)}$ is to be a solution of the matrix instance $\underline{\mathrm{X}}^{(k)}$. This is best explained by imagining that Arthur is trying to find the orthogonal matrix and Merlin already knows the orthogonal matrix but has still been following the steps performed so far. Recall that we are now at a point where

$$\begin{aligned} \sum a'(x)x &= \left\langle w\right| X_h' \left|w\right\rangle - \left\langle v\right| X_g' \left|v\right\rangle \\ &= \left\langle w\right| X_h' \left|w\right\rangle - \left\langle w\right| O X_g' O^T \left|w\right\rangle \\ &= 0. \end{aligned}$$

From Merlin's point of view along the $|w\rangle$ direction the ellipsoids X'_h and OX'_gO^T touch. Suppose he started with the ellipsoids X'_h, X'_g and only subsequently rotated the second one. He can mark the point along the direction $|v\rangle$ on the X'_g ellipsoid as the point that would after rotation touch the X'_h ellipsoid because as $X'_g \to OX'_gO^T$ the point along the $|v\rangle$ direction would get mapped to the point along the direction $O|v\rangle = |w\rangle$. Now, since the ellipsoids touch it must be so, Merlin deduces, that the normal of the ellipsoid X'_g at the point $|v\rangle / \sqrt{\langle v| X'_g |v\rangle}$ is mapped to the normal of the ellipsoid X'_h at the point $|w\rangle / \sqrt{\langle w| X'_h |w\rangle}$ when X'_g is rotated to OX'_gO^T, i.e. $O|u_g\rangle = |u_h\rangle$.

From Arthur's point of view, who has been following Merlin's reasoning, in addition to knowing that O must satisfy $O|v\rangle = |w\rangle$ he now knows that it must also satisfy $O|u_g\rangle = |u_h\rangle$.

Merlin further concludes that the curvature of the X'_g ellipsoid at the point $|v\rangle / \sqrt{\langle v| X'_g |v\rangle}$ must be more than the curvature of the X'_h ellipsoid at the point $|w\rangle / \sqrt{\langle w| X'_h |w\rangle}$. To be precise, he needs to find a method for evaluating this curvature. He knows that the brute-force way of doing this is to find a coordinate system with its origin on the said point and then imagining the manifold, locally, as a function from $n-1$ coordinates to one coordinate, call it $x_n(x_1, x_2 \dots x_{n-1})$ (think of a sphere centred at the origin; it can be thought of, locally, as a function from x and y to z given by $z = \sqrt{x^2+y^2}$). The curvature of this object is a generalisation of the second derivative which forms a matrix with its elements given by $\partial_{x_i}\partial_{x_j}x_n$. Since this matrix is symmetric he knows it can be diagonalised. The directions of the eigenvectors of this matrix he calls the principle directions of curvature where the curvature values are the corresponding eigenvalues. He recalls that there is a simpler way of evaluating these principle directions and curvatures which uses the Weingarten map. The eigenvectors of the Reverse Weingarten map W'_h, evaluated for X'_h at $|w\rangle$, yield the normal and tangent vectors with the corresponding eigenvalues zero and radii of curvature respectively. Curvature is the inverse of the radius of curvature. Similarly for the Reverse Weingarten map W'_g evaluated for X'_g at $|v\rangle$.

With this knowledge Merlin deduces that he can write, for some $\tilde{O}_{ij} \in \mathbb{R}$ such that $\sum_j \tilde{O}_{ij}\tilde{O}_{jk} = \delta_{ik}$,

$$
\begin{aligned}
O^{(k)} &= |u_h\rangle\langle u_g| + \sum_{i,j} \tilde{O}_{ij} |t_{h_i}\rangle\langle t_{g_j}| \\
&= \left(|u_h\rangle\langle u_h| + \underbrace{\sum_{i,j} \tilde{O}_{ij} |t_{h_i}\rangle\langle t_{h_j}|}_{=O^{(k-1)}} \right) \left(\underbrace{|u_h\rangle\langle u_g| + \sum_i |t_{h_i}\rangle\langle t_{g_i}|}_{=\bar{O}^{(k)}} \right)
\end{aligned}
$$

where he re-introduced the superscript in the orthogonal operators. He then turns to his intuition about the curvature of the smaller ellipsoid being more than that of the larger ellipsoid. He observes that equivalently, the radius of curvature of the smaller ellipsoid must be smaller than that of the larger ellipsoid. To make this precise he first notes that the Weingarten map W'_g gets transformed to OW'_gO^T when X'_g is rotated as OX'_gO^T. He considers the point $|w\rangle / \sqrt{\langle w| X'_h |w\rangle}$, which is shared by both the X'_h and the OX'_gO^T ellipsoid. It must be so, he reasons, that along all directions in the tangent plane, the X'_h ellipsoid (the smaller one, remember larger X'_h means smaller ellipsoid) must have a smaller radius of curvature than the OX'_gO^T ellipsoid, i.e. for all

$|t\rangle \in \text{span}\{|t_{h_i}\rangle\}$, $\langle t| W'_h |t\rangle \leq \langle t| OW'_g O^T |t\rangle$. Restricting his attention to the tangent space he deduces the statement is equivalent to $W'_h \leq OW'_g O^T$. He writes this out explicitly as $\sum c_{h_i}^{-1} |t_{h_i}\rangle \langle t_{h_i}| \leq \sum c_{g_i}^{-1} O |t_{g_i}\rangle \langle t_{g_i}| O^T$. Now he uses the form of O he had deduced to obtain $\sum c_{h_i}^{-1} |t_{h_i}\rangle \langle t_{h_i}| \leq \sum c_{g_i}^{-1} O^{(k-1)} |t_{h_i}\rangle \langle t_{h_i}| O^{(k-1)T}$. From this he is able to deduce that the inequality $X_h^{(k-1)} \geq O^{(k-1)} X_g^{(k-1)} O^{(k-1)T}$ must hold.

Merlin's reasoning entails, Arthur summarises, that $O^{(k)}$ must always have the form

$$O^{(k)} = \left(\left| u_h^{(k)} \right\rangle \left\langle u_h^{(k)} \right| + O^{(k-1)} \right) \bar{O}^{(k)}$$

and that $O^{(k-1)}$ must satisfy the constraint

$$X_h^{(k-1)} \geq O^{(k-1)} X_g^{(k-1)} O^{(k-1)T}.$$

Merlin, surprised by the similarity of the constraint he obtained with the one he started with, extends his reasoning to the vector itself. He knows that $O^{(k)} \left| v^{(k)} \right\rangle = \left| w^{(k)} \right\rangle$ but now he substitutes for $O^{(k)}$ to obtain $\left(\left| u_h^{(k)} \right\rangle \left\langle u_h^{(k)} \right| + O^{(k-1)} \right) \bar{O}^{(k)} \left| v^{(k)} \right\rangle = \left| w^{(k)} \right\rangle$. He observes that $O^{(k-1)}$ can not influence the $\left| u_h^{(k)} \right\rangle$ component of the vector $\bar{O}^{(k)} \left| v^{(k)} \right\rangle$. He thus projects out the $\left| u_h^{(k)} \right\rangle$ component to obtain

$$O^{(k-1)} \underbrace{\left(\bar{O}^{(k)} \left| v^{(k)} \right\rangle - \langle u_h | \bar{O}^{(k)} \left| v^{(k)} \right\rangle |u_h\rangle \right)}_{=\left| v^{(k-1)} \right\rangle} = \underbrace{\left| w^{(k)} \right\rangle - \left\langle u_h^{(k)} | w^{(k)} \right\rangle \left| u_h^{(k)} \right\rangle}_{=\left| w^{(k-1)} \right\rangle}.$$

With this, Arthur realises, he can reduce his problem involving a k-dimensional orthogonal matrix into a smaller problem in $k-1$ dimensions with exactly the same form. Since Merlin's orthogonal matrix was any arbitrary solution, and since the constraints involved do not depend explicitly on the solution (only on the initial problem), Arthur concludes that this reduction must hold for all possible solutions.

– **Wiggle-v Method: If $\lambda = -\xi^{(k)}$ or $\lambda = -\chi^{(k)}$ then**
The aforesaid method relies on matching the normals. It works well so long as the correct operator monotone (the monotone that yields X'_h and X'_g for which $|w\rangle / \sqrt{\langle w| X'_h |w\rangle}$ is a point on both X'_h and $OX'_g O^T$) doesn't yield infinities. If the operator monotone yields infinities it means that one of the directions involved has infinite curvature which in turn means that the component of the normal along this direction can be arbitrary. To see this, imagine having a line contained inside an ellipsoid (both centred at the origin) touching its boundaries. The line can be thought of as an ellipse with infinite curvature along one of the directions. The normal of the line at its tip is arbitrary and therefore we can't require the usual condition that normals of the two curves must coincide. The solution is to consider the sequence leading to the aforesaid situation.

1. $\left| u_h^{(k)} \right\rangle$ is renamed to $\left| \bar{u}_h^{(k)} \right\rangle$, $\left| u_g^{(k)} \right\rangle$ remains the same.
2. Let $\tau = \cos\theta := \left\langle u_g^{(k)} | v^{(k)} \right\rangle / \left\langle \bar{u}_h^{(k)} | w^{(k)} \right\rangle$. Let $\left| \bar{t}_h^{(k)} \right\rangle$ be an eigenvector of $X_h'^{(k)-1}$ with zero eigenvalue (comment: this is also perpendicular to $\left| w^{(k)} \right\rangle$). Redefine

$$\left| u_h^{(k)} \right\rangle := \cos\theta \left| \bar{u}_h^{(k)} \right\rangle + \sin\theta \left| \bar{t}_h^{(k)} \right\rangle,$$

$$\left| t_{h_k}^{(k)} \right\rangle = s \left(-\sin\theta \left| \bar{u}_h^{(k)} \right\rangle + \cos\theta \left| \bar{t}_h^{(k)} \right\rangle \right)$$

where the sign $s \in \{1, -1\}$ is fixed by demanding $\left\langle t_{h_k}^{(k)} | w^{(k)} \right\rangle \geq 0$.

3. $\bar{O}^{(k)}$ and $\left|v^{(k-1)}\right\rangle, \left|w^{(k-1)}\right\rangle$ are evaluated as step 1 and 2 of the finite case.
4. Define

$$X_h'^{(k-1)} := \mathrm{diag}\{c_{h_1}^{(k)}, c_{h_2}^{(k)}, \ldots, c_{h_{k-1}}^{(k)}\}, \qquad X_g'^{(k-1)} := \mathrm{diag}\{c_{g_1}^{(k)}, c_{g_2}^{(k)}, \ldots, c_{g_{k-1}}^{(k)}\}.$$

Let $[\chi'^{(k-1)}, \xi'^{(k-1)}]$ denote the smallest interval containing $\mathrm{spec}[X_h'^{(k-1)} \oplus X_g'^{(k-1)}]$. Let $\lambda' = -\chi'^{(k-1)} + 1$ where instead of 1 any positive number would also work. Consider $f_{\lambda''}$ on $[\chi'^{(k-1)}, \xi'^{(k-1)}]$. Let $\eta = -f_{\lambda'}(\chi'^{(k-1)}) + 1$. Define

$$X_h^{(k-1)} := f_{\lambda'}(X_h'^{(k-1)}) + \eta, \qquad X_g^{(k-1)} := f_{\lambda'}(X_g'^{(k-1)}) + \eta.$$

5. **Jump** to **End**.

We start with the case $\lambda = -\xi^{(k)}$. The other case with $\lambda = -\chi^{(k)}$ follows analogously. For the moment just imagine $\eta = 0$ for simplicity; for $\eta \neq 0$ the argument goes through essentially unchanged. Note that because $\langle w| f_{-\xi}(X_h) |w\rangle - \langle v| f_{-\xi}(X_g) |v\rangle$ is zero we can conclude that $y_{h_i}^{(k)} = \xi$ implies $q_{h_i} = 0$. After the application of the map $f_{-\xi}$ these $y_{h_i}^{(k)}$s and $y_{g_i}^{(k)}$s would become infinities but $\left\langle t_{h_i}^{(k+1)} | w \right\rangle$ and $\left\langle t_{g_i}^{(k+1)} | v \right\rangle$ would be zero where we suppressed the superscripts for $\left|v^{(k)}\right\rangle$ and $\left|w^{(k)}\right\rangle$. Since the eigenvalues are arranged in the ascending order in $X_h^{(k)}$ (in the $\left\{\left|t_{h_i}^{(k+1)}\right\rangle\right\}$ basis) we have $y_{h_k}^{(k)} = \xi$ and the corresponding vector is $\left|t_{h_k}^{(k+1)}\right\rangle =: |\bar{t}_h\rangle$. It would be useful to define $\left|\tilde{t}_{h_i}\right\rangle = |t_{h_i}\rangle$ for $i = 1, 2, \ldots j-1$ and $|\bar{t}_{h_l}\rangle = |t_{h_i}\rangle$ for $i = j, j+1, \ldots k$, $l = (i-j)+1$ where j is the smallest i for which $x_{h_i} = \xi$ (their existence is a straight forward consequence of dimension counting, $k \geq n_g + n_h - 1$). This allows us to speak of the subspace with eigenvalue ξ of $X_h^{(k)}$ easily. We focus on the two dimensional plane spanned by $|w\rangle$ and $|\bar{t}_h\rangle$.

Consider the M-view (Merlin's point of view). Since Merlin has a solution $O^{(k)}$ to the matrix instance

$$\underline{\mathrm{X}}^{(k)} = \{X_h^{(k)}, X_g^{(k)}, \left|w^{(k)}\right\rangle, \left|v^{(k)}\right\rangle\}$$

his solution is also a solution to the matrix instance

$$\underline{\mathrm{X}}^{(k)}(\lambda) := \left\{f_\lambda(X_h^{(k)}), f_\lambda(X_g^{(k)}), \left|w^{(k)}\right\rangle, \left|v^{(k)}\right\rangle\right\}$$

for $\lambda \leq -\xi$ but close enough to $-\xi$ such that $f_\lambda(X_h), f_\lambda(X_g) > 0$. This is a consequence of f_λ being operator monotone. Using Corollary 138 and Lemma 139 we know that since the ellipsoids corresponding to the matrix instance $\underline{\mathrm{X}}(-\xi)$ touch along $|w\rangle$ (as we are given that $\langle w| f_{-\xi}(X_h) |w\rangle - \langle w| O f_{-\xi}(X_g) O^T |w\rangle = \langle w| f_{-\xi}(X_h) |w\rangle - \langle v| f_{-\xi}(X_g) |v\rangle = 0$) there must also exist some vector $|c(\lambda)\rangle$ such that $\langle c(\lambda)| f_\lambda(X_h) |c(\lambda)\rangle - \langle c(\lambda)| O f_\lambda(X_g) O^T |c(\lambda)\rangle = 0$ that is the ellipsoids corresponding to the matrix instance $\underline{\mathrm{X}}(\lambda)$ touch along the said direction. (Caution: Do not confuse $|c(\lambda)\rangle$ with c_{h_i}/c_{g_i}. The latter are used for curvature values and the former refers to the contact vector just defined.) Note that to match the other conditions of the lemma it suffices to assume that X_h and X_g do not have a common eigenvalue which in turn is guaranteed by the "remove spectral collision" part.

It is easy to convince oneself that $\lim_{\lambda \to -\xi} |c(\lambda)\rangle = |w\rangle$ (hint: argue along the lines $f_\lambda(X_h)$ is very close to $f_{-\xi}(X_h)$ and so the vectors should also be very close which satisfy the condition). Note that we can write

$$|w\rangle = \sum_{i=1}^{j-1} q_{h_i} \left|\tilde{t}_{h_i}\right\rangle$$

because $\langle \bar{t}_{h_i} | w \rangle = 0$. There is no such restriction on $|c(\lambda)\rangle$ which can have the more general form $|c(\lambda)\rangle = \sum_{i=1}^{j-1} c(\lambda)_i \left|\tilde{t}_{h_i}\right\rangle + \sum_{i=j}^{k} c(\lambda)_i \left|\bar{t}_{h_l}\right\rangle$ where $l = (i-j)+1$. Restating one of the limit conditions, for $i = j, j+1 \ldots k$, we must have the $\lim_{\lambda \to -\xi} c(\lambda)_i = 0$. At this point we use the fact that if O is a solution it entails that

$$\acute{O}(\lambda) := \left(\sum_{i=1}^{j-1} \left|\tilde{t}_{h_i}\right\rangle \left\langle \tilde{t}_{h_i}\right| + \sum_{i,m=1}^{k-j+1} Q(\lambda)_{im} \left|\bar{t}_{h_i}\right\rangle \left\langle \bar{t}_{h_m}\right| \right) O$$

is also a solution, where $Q(\lambda)$ is an orthogonal matrix in the space spanned by $\{|\bar{t}_{h_i}\rangle\}$. This is a consequence of the fact that $\{|\bar{t}_{h_i}\rangle\}$ spans an eigenspace (with the same eigenvalue, $f_\lambda(\xi)$,) of $f_\lambda(X_h)$. We can use this freedom to ensure that the point of contact always has the form

$$|c(\lambda)\rangle = \sum_{i=1}^{j-1} c(\lambda)_i \left|\tilde{t}_{h_i}\right\rangle + \bar{c}(\lambda) \left|\bar{t}_h\right\rangle$$

where $\bar{c}(\lambda) = \sqrt{\sum_{i=j}^{k} c(\lambda)_i^2}$ which must vanish in the limit $\lambda \to -\xi$ as its constituents disappear in the said limit. Similarly $\lim_{\lambda \to -\xi} c(\lambda)_i = q_{h_i}$.

Next we evaluate the normals $|u_h(\lambda)\rangle$ at $|c(\lambda)\rangle$ for the ellipsoid represented by $f_\lambda(X_h)$ and similarly the normal $|\bar{u}_h\rangle$ at $|w\rangle$ for the ellipsoid represented by $f_{-\xi}(X_h)$ to show that $\lim_{\lambda \to -\xi} |u_h(\lambda)\rangle \neq |\bar{u}_h\rangle$ (see Figure 6.3). Notice that the right-most term in $|u_h(\lambda)\rangle = \mathcal{N}\left[\sum_{i=1}^{j-1} f_\lambda(y_{h_i}) c(\lambda)_i \left|\tilde{t}_{h_i}\right\rangle + f_\lambda(\xi)\bar{c}(\lambda)\left|\bar{t}_h\right\rangle\right]$ has $f_\lambda(\xi)$ approaching infinity and $\bar{c}(\lambda)$ approaching zero as λ tends to $-\xi$. This is why it can have a finite component along $|\bar{t}_h\rangle$. On the other hand, $|\bar{u}_h\rangle = \mathcal{N}\left[\sum_{i=1}^{j-1} f_{-\xi}(y_{h_i}) q_{h_i} \left|\tilde{t}_{h_i}\right\rangle\right]$ which has no component along $|\bar{t}_h\rangle$. Since $\lim_{\lambda \to -\xi} f_\lambda(y_{h_i}) = f_{-\xi}(y_{h_i})$ and $\lim_{\lambda \to -\xi} c(\lambda)_i = q_{h_i}$ for $i \in \{1, 2 \ldots j-1\}$, we can write

$$\lim_{\lambda \to -\xi} |u_h(\lambda)\rangle = \cos\theta \left|\bar{u}_h\right\rangle + \sin\theta \left|\bar{t}_h\right\rangle := |u_h\rangle .$$

Evidently, we must use $|u_h\rangle$ instead of $|\bar{u}_h\rangle$ to be able to use the reasoning of the finite method. However, we do not know $\cos\theta$ yet.

Our strategy is to proceed as in the finite method with the assumption that $|c(\lambda)\rangle$ is known (which it isn't as we only know it exists and how it behaves in the limit of $\lambda \to -\xi$) and then use a consistency condition to find $\cos\theta$ in terms of known quantities. At this point we re-introduce the superscripts as we will reduce the dimension of the problem as we proceed. Let the normal and tangents at $O^T |c(\lambda)\rangle$ for $f_\lambda(X_g)$ be given by $\left\{\left|u_g^{(k)}(\lambda)\right\rangle, \left\{t_{g_i}^{(k)}(\lambda)\right\}\right\}$. Similarly at $|c(\lambda)\rangle$ for $f_\lambda(X_h)$ the normal and tangents are $\left\{\left|u_h^{(k)}(\lambda)\right\rangle, \left\{t_{h_i}^{(k)}(\lambda)\right\}\right\}$. From the finite method we know that $O^{(k)}(\lambda) := \left(|u_h(\lambda)\rangle\langle u_h(\lambda)| + O^{(k-1)}\right)\bar{O}^{(k)}$ where $\bar{O}^{(k)} = \left|u_h^{(k)}(\lambda)\right\rangle\left\langle u_g^{(k)}(\lambda)\right| + \sum_i \left|t_{h_i}^{(k)}\right\rangle\left\langle t_{g_i}^{(k)}\right|$ can be used to reduce the problem into a smaller instance of itself. In particular, we must have $\left\langle u_h^{(k)}(\lambda)|w\right\rangle = \left\langle u_h^{(k)}(\lambda)\right| O^{(k)}(\lambda)\,|v\rangle = \left\langle u_g^{(k)}(\lambda)|v\right\rangle$ because $O^{(k-1)}$ can influence only the subspace spanned by $\left\{\left|t_{h_i}^{(k)}\right\rangle\right\}$ and the component of the vectors $|w\rangle$ and $O^{(k)}|v\rangle$ along $\left|u_h^{(k)}(\lambda)\right\rangle$ must match for consistency.

We can determine $\cos\theta$ by taking the limit of the aforesaid condition as $\langle u_h|w\rangle = \langle u_g|v\rangle$ where we again suppressed the superscripts. Substituting $|u_h\rangle = \cos\theta\,|\bar{u}_h\rangle +$

$\sin\theta\,|\bar{t}_h\rangle$ we obtain

$$\cos\theta = \frac{\langle u_g|v\rangle}{\langle \bar{u}_h|w\rangle}.$$

It now remains to find the limit of the reverse Weingarten maps. The reverse Weingarten map for $f_\lambda(X_g)$ along the normal $|u_g(\lambda)\rangle$ is not of concern because it has a well defined limit as $\lambda \to -\xi$. We consider the case for $f_\lambda(X_h)$ along the normal $|u_h(\lambda)\rangle$. Note that the support function as defined in Equation (6.1) is finite in the limit $\lambda \to -\xi$ (use the definition of the normal to get $\sum x_i^{-1}u_i^2 = \sum x_i^{-1}x_i^2c_i^2 = \sum x_i c_i^2 = \langle c|\,X\,|c\rangle$, plug in $|c\rangle = |w\rangle$, $X = f_{-\xi}(X_h)$ and then use the fact that $\langle w|\,f_{-\xi}(X_h)\,|w\rangle - \langle v|\,f_{-\xi}(X_g)\,|v\rangle = 0$ which means both must be finite by noting that we already dealt with the troublesome case of $\infty - \infty$ in the "remove spectral collision" part). Let us denote it by $h(\lambda)$. Now the reverse Weingarten map as defined in Equation (6.2) is given by

$$(W_h(\lambda))_{im} = -\frac{1}{h(\lambda)^2}\frac{u_{h_i}(\lambda)u_{h_m}(\lambda)}{f_\lambda(y_{h_i})f_\lambda(y_{h_m})} + \frac{\delta_{im}}{f_\lambda(x_{h_i})}.$$

Since $\lim_{\lambda\to-\xi}|u(\lambda)\rangle$ is well defined, $\lim_{\lambda\to-\xi}h(\lambda)$ is finite, we only need to show that $\lim_{\lambda\to-\xi}1/f_\lambda(y_{h_i})$ is well defined. (We assumed η is zero so $f_{-\xi}(y_{h_i}) \neq 0$. If η is not zero we must consider $f_{-\xi}(y_{h_i}) + \eta$ everywhere but that changes no argument.) For $i = 1, 2 \ldots j-1$, $f_{-\xi}(y_{h_i})$ is finite but for $i = j, j+1 \ldots k$, $f_{-\xi}(y_{h_i})$ is not well defined however $1/f_{-\xi}(y_{h_i}) = 0$. We therefore conclude that

$$\lim_{\lambda\to-\xi}(W_h(\lambda))_{im} = \begin{cases} -\frac{1}{h^2}\frac{u_{h_i}u_{h_m}}{f_{-\xi}(y_{h_i})f_{-\xi}(y_{h_m})} + \frac{\delta_{im}}{f_{-\xi}(y_{h_i})} & i,m \in \{1,2\ldots j-1\} \\ 0 & i,m \in \{j, j+1\ldots k\} \end{cases} := (W_h)_{im}$$

which is simply the reverse Weingarten map evaluated for $f_{-\xi}(X_h)$ along $|u_h\rangle = \cos\theta\,|\bar{u}_h\rangle + \sin\theta\,|\bar{t}_h\rangle$ and $\cos\theta = \langle u_g|v\rangle\,/\,\langle\bar{u}_h|w\rangle$. It remains to relate W_h with the reverse Weingarten map, $\bar{W}_h$, evaluated for $f_{-\xi}(X_h)$ along $|\bar{u}_h\rangle$. Surprisingly, it is easy to see that $W_h = \bar{W}_h$ because only the $\cos\theta\,|\bar{u}_h\rangle$ part contributes to the non-zero portion of W_h and the $\cos\theta$ factor gets cancelled due to the h^2 term. Further, recall that the normal vector is always an eigenvector of the reverse Weingarten map evaluated along it, with eigenvalue zero. This tells us that if there is(are) tangent(s) with zero radius of curvature then the normal is not uniquely defined. This confirms what we already knew. Now since both $|\bar{u}_h\rangle, |\bar{t}_h\rangle$ have zero eigenvalues for $\bar{W}_h(= W_h)$ and $|u\rangle = \cos\theta\,|\bar{u}_h\rangle + \sin\theta\,|\bar{t}_h\rangle$ we define $|t_h\rangle := s\,(\sin\theta\,|\bar{u}_h\rangle - \cos\theta\,|\bar{t}_h\rangle)$ to span the same space so that $|u\rangle$ is the correct normal vector (as we deduced earlier in our discussion) and $|t_h\rangle$ is the correct tangent vector corresponding to the point $|w\rangle$ of $f_{-\xi}(X_h)$.

The final step is to convert the condition on the reverse Weingarten map into a condition on the Weingarten map (inverse of the reverse Weingarten map). After extracting the tangent vectors appropriately, one simply needs to add a constant before inverting to obtain the Weingarten map condition. This is done in the last step. This completes the proof of the wiggle-v method for $\lambda = -\xi$.

To see how the same reasoning applies to the $\lambda = -\chi$ case first note that for $\lambda \geq -\chi$ we have $f_\lambda(X_h), f_\lambda(X_g) < 0$ (assuming $\eta = 0$ as before). The condition $f_\lambda(X_h) \geq Of_\lambda(X_g)O^T$ can then be expressed as $-f_\lambda(X_g) \geq -O^T f_\lambda(X_h)O$ with $O^T\,|w\rangle = |v\rangle$ which can now be reasoned analogous to the aforementioned analysis.

– **End**: Restart the current phase (phase 2) with the newly obtained $(k-1)$ sized objects. We end with giving the dimension argument. The dimension after every iteration is $k-1 \geq n_g^{(k-1)} + n_h^{(k-1)} - 1$ if we start with the assumption that $k \geq n_g^{(k)} + n_h^{(k)} - 1$. The reason is that either $n_g^{(k-1)} = n_g^{(k)} - 1$ or $= n_g^{(k)}$. Similarly, either $n_h^{(k-1)} = n_h^{(k)} - 1$ or $= n_h^{(k)}$. Justification of this is simply that we remove at least one component from the two vectors (from the $n_g^{(k)}$ for the usual wiggle-v). To see this, note that in the finite case we remove one from both as we write express the vector in a new basis. This new basis is the space where the vector has finite support. We then remove one of the components in the sub-problem. In the infinite case, it is possible that we remove one and add one for $n_h^{(k-1)}$, assuming it is the usual wiggle-v, but we necessarily reduce $n_g^{(k-1)}$ as this is similar to the finite case. For the other wiggle-v, g and h get swapped but the counting stays the same.

6.3.3.3 Phase 3: Reconstruction

Let k_0 be the iteration at which the algorithm stops. Using the relation

$$O^{(k)} = \bar{O}_g^{(k)} \left(\left| u_h^{(k)} \right\rangle \left\langle u_h^{(k)} \right| + O^{(k-1)} \right) \bar{O}_h^{(k)}$$

(or its transpose if $s^{(k)} = -1$), evaluate $O^{(k_1)}$ from $O^{(k_0)} := \mathbb{I}_{k_0}$, then $O^{(k_2)}$ from $O^{(k_1)}$, then $O^{(k_3)}$ from $O^{(k_2)}$ and so on until $O^{(n)}$ is obtained which solves the matrix instance $\underline{\mathrm{X}}^{(n)}$ we started with. In terms of EBRM matrices, the solution is given by $H = X_h^{(n)}$, $G = O^{(n)} X_g O^{(n)T}$, and $|w\rangle = \left|w^{(n)}\right\rangle$.

§ 6.4 Conclusion

In this chapter, we described the EMA algorithm which allows us to numerically find the unitaries corresponding to arbitrary Λ-valid functions (moves used in point games), which combined with the framework, allows us to numerically find quantum WCF protocols corresponding to any TIPG, including Mochon's TIPG and the Pelchat-Høyer TIPGs both of which achieve arbitrarily small bias. A preliminary implementation of the EMA algorithm on python [5], which is usable but not automated enough for an end-user, yielded the following results. Note that, chronologically, these results were obtained before those of Chapter 5.

1. Mochon's denominator needs neither padding nor operator monotones. For assignments given by Mochon's denominator (see Lemma 76), we saw that $\langle x_h \rangle = \langle x_g \rangle$ which means that for the first iteration of the algorithm, we need not use any operator monotone function. This was clear. What was surprising at the time was that even for subsequent iterations, one need not use operator monotones, which also explained why[2] we did not need padding, i.e. the (solution) orthogonal matrix had size $n \times n$ for $n = n_g = n_h$ (and not more). In Chapter 7 we prove this analytically and follow this approach to construct a more general solution geometrically which covers Mochon's assignments. We later realised that this solution can even be stated without using ellipsoids which comprised Chapter 5.

2. *Moves in the bias* $1/18$ *protocol do not need padding (no wiggle-v).* We already know analytically that there are specific cases where padding is required. However, when we tried to numerically implement the moves involved in Mochon's protocols going as low as $\epsilon = 1/18$, to our surprise, we found that in no case was padding necessary (which means the wiggle-v method was never invoked). It would be interesting to see if this can be proven to be the case for all of Mochon's moves. There is another class of games that achieve arbitrarily low bias due to Pelchat and Høyer [21] whose moves are also worth investigating in this regard, and even otherwise.

3. *Trick to improve the precision of the EMA algorithm.* The algorithm tries to find a λ such that $\langle w| f_\lambda(X_h) |w\rangle - \langle v| f_\lambda(X_g) |v\rangle = 0$. In the finite case, one must also have for consistency, $\langle w|n_h(\lambda)\rangle = \langle v|n_g(\lambda)\rangle$ (this is because in subsequent steps, the space orthogonal is affected so if the component of the honest states along the normals is not mapped correctly, it would not get fixed later; this would mean there is no solution as we are only imposing necessary conditions). We observed that, numerically, we get a better precision if we use the latter condition for fine-tuning the result (after applying the former for obtaining a more course-grained solution). While analytically, the first condition implies the latter exactly, this ceases to be the case numerically due to the finiteness of precision. A careful error analysis of the EMA algorithm would be required to fully understand this behaviour. As a first step, we can understand this improvement as a direct consequence of the fact that the honest state is explicitly mapped correctly (up to the precision of the machine, which is about 16 floating points for most computers) if we use the method involving normals while in the latter, this should happen implicitly.

Limitations of the current implementation.

1. *Limited wiggle-v.* We have not fully implemented the Wiggle-v method which means that it would be cumbersome to apply it to the general merge and split, for instance. However, for them we already give the explicit Blinkered Unitaries. For the rest, as we already saw, it does not even seem necessary.

2. *Minor pending issues.* Sometimes due to noise (arising from finiteness of the precision) our global minimiser gets trapped into local minima and has to be guided manually by looking at the graph. This means that a refined algorithm should also be able to solve the problem. Further, we did

[2] To see this, note that the only time we spill over to the extra dimensions, is when we use the wiggle-v method. Otherwise, we stay inside the first $\max(n_g, n_h)$ dimensions.

not implement the systematic method defined by the EMA algorithm for finding the spectrum of the matrices it uses but it appears that almost any guess works for Mochon's assignments.

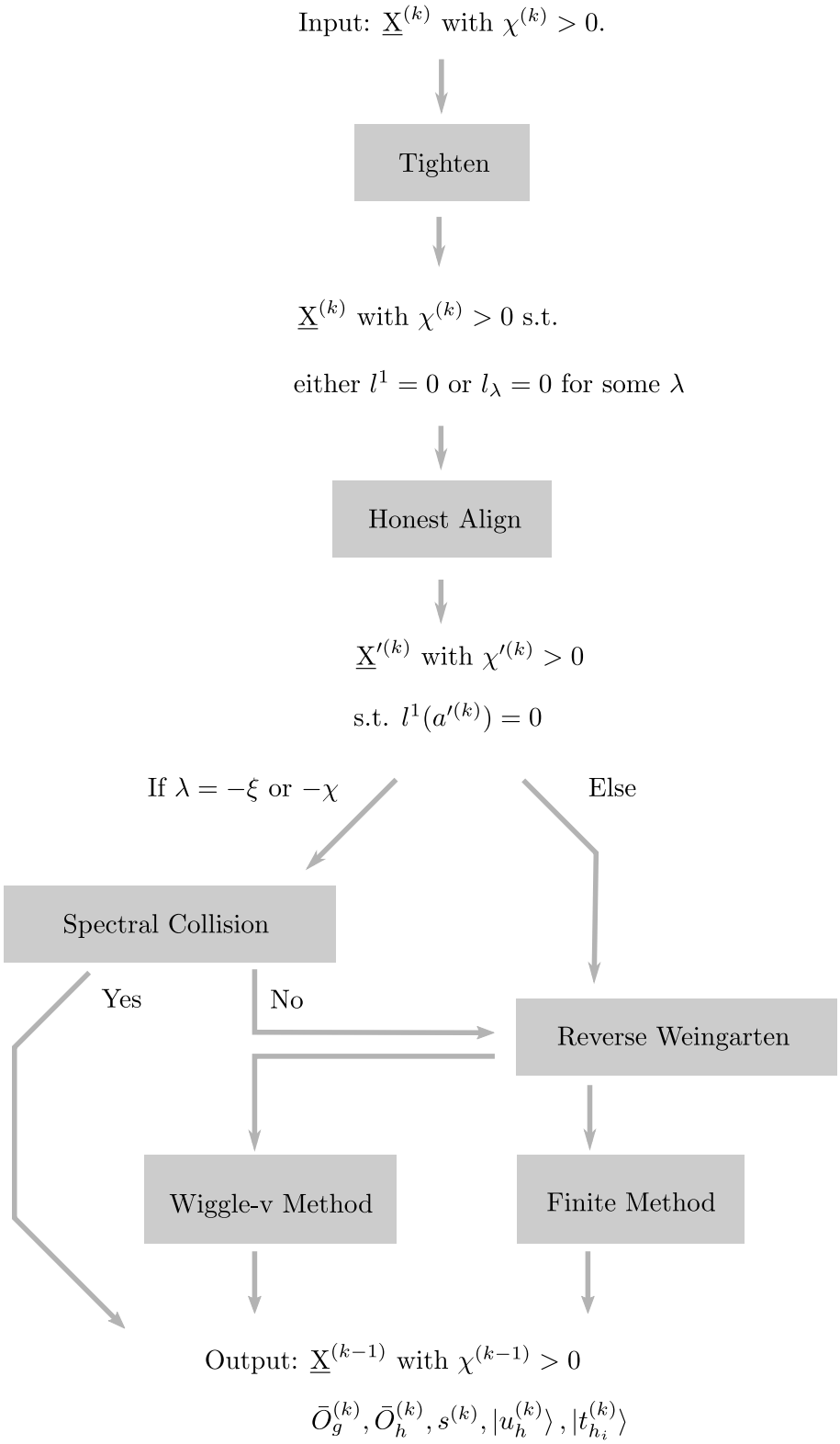


Figure 6.2: Overview of the main step, the iteration, of the algorithm (excluding the boundary condition).

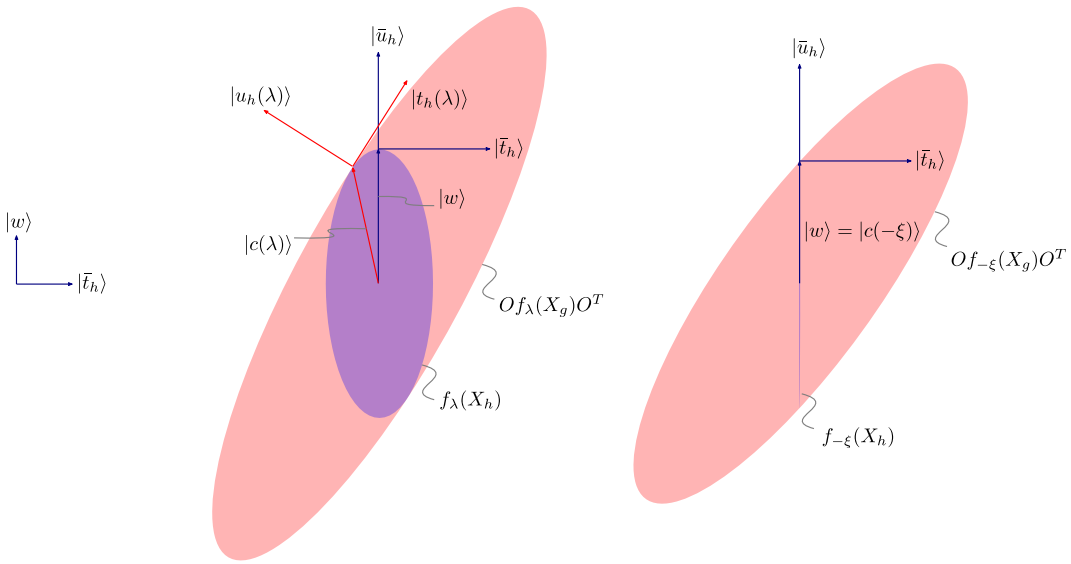


Figure 6.3: A sequence leading to infinite curvature.

7

Chapter

Approaching bias $1/(4k+2)$ | a geometric approach

We outlined an argument at the end of Theorem 75 (Chapter 4), and filled in the details in Section 6.1 (Chapter 6), to show that one can restrict oneself to real matrices without loss of generality for the discussion of weak coin flipping protocols[1]. In Section 6.2 (again Chapter 6), we saw how this allows one to interpret the matrices as ellipsoids and the inequalities as containment thereof. Further, it allowed us to use tools and intuition from geometrical analysis. The appeal of this approach was that it allowed us to find the unitaries (in fact orthogonal matrices) corresponding to any valid move. The limitation was that the solution was only numerical.

We also saw, in Chapter 5, how one can construct an exact solution to one class of valid moves, namely Mochon's assignments. This was done in two steps, first we established that Mochon's assignment can be expressed as a sum of either monomial or effectively monomial assignments. Subsequently, we gave the exact unitaries/orthogonal matrices for monomial assignments.

Our goal in this chapter, is to construct an analytic solution to a monomial assignment, using the geometric approach. In fact, we first found an exact solution using this approach and later found a way of bypassing the ellipsoid machinery, as described in Chapter 5. Nonetheless, we hope that having multiple ways of seeing the same problem contributes towards constructing a general analytic solution (i.e. solution to all valid functions).

To construct an analytic solution (for monomial assignments, using the geometric approach), we introduce some notation which overlaps partially with that of Chapter 5 but diverges as it is built further. An attempt has been made to keep this chapter accessible to the readers who have skipped Chapter 6, at the cost of slight redundancy.

Notation. We use $(a,b,c) \oplus (d,e) = (a,b,c,d,e)$. Suppose S is a 4-tuple (an ordered list with 4 elements) and we wish to refer to the third element of S. We write this as

$$(*,*,p,*) := S. \tag{7.1}$$

In the interest of conciseness, a matrix of rank at most k is denoted by $M^{\bar{k}}$. We always use a bar in the superscript to distinguish it from powers. For instance, $(M^{\bar{k}})^2$ refers to the square of the rank k matrix $M^{\bar{k}}$.

§ 7.1 Ellipsoid Picture

In this section we revisit the ellipsoid picture which differs from the discussion in Chapter 6 in two specific ways.

First, recall that in the EMA algorithm, to consider sub-instances of the problem, we stated the sub-problem in an appropriately reduced subspace, i.e. the tangent space of the associated initial ellipsoids.

[1] (granted one is only concerned about lowering the bias and not other parameters such as resource requirements)

Consequently, the *dimensions* of the matrices describing the sub-problems dropped as the algorithm proceeded. Here, we take a slightly different approach. We always consider matrices of the same dimension but every sub-problem is described by matrices of one *rank* less than its parent problem.

Second, for the EMA algorithm, we were targetting a numerical solution. Consequently, finding inverses of Hermitian matrices was taken for granted. Here, we are targetting an exact solution. We therefore both track the dependence on inverses and give methods of evaluating inverses of the matrices we subsequently use.

7.1.1 Exact Formulae

Given a positive matrix we can associate an ellipsoid with it. We introduce projectors from the outset to handle low rank matrices which become rife in the analysis later.

Notation. Given a projector Π, we denote the set $\{\Pi |v\rangle \,|\, |v\rangle \in \mathbb{R}^n\}$ by $\Pi\mathbb{R}^n$.

Definition 145 (Ellipsoid and Map)**.** Given an $n \times n$ matrix $G \geq 0$, let Π be a projector onto the non-zero eigenvalue eigenspace of G. The *ellipsoid* (or more precisely the *Ellipsoidal Manifold*) associated with G is given by $S_G := \{|s\rangle \in \Pi\mathbb{R}^n | \langle s| G |s\rangle = 1\}$. The *Ellipsoid Map*, $\mathcal{E}_G : \Pi\mathbb{R}^n \to \Pi\mathbb{R}^n$, is defined as $\mathcal{E}_G(|v\rangle) = |v\rangle / \sqrt{\langle v| G |v\rangle}$.

Note that $\langle s| G |s\rangle = 1$ is essentially of the form $\sum_i g_i s_i^2 = 1$ where $G = \sum_i g_i |i\rangle \langle i|$ and $|s\rangle = \sum_i s_i |i\rangle$, i.e. the equation of an ellipsoid, justifying our choice of words. Recall that we had introduced the use of the turnstile symbol ($\dashv$) to represent the inverse of a matrix $G \geq 0$ on its non-zero eigenspace. We restate it for convenience.

Definition 146 (Positive Inverse)**.** Given a symmetric matrix $G \geq 0$, let $\Pi^\perp = \mathbb{I} - \Pi$ be the projector onto its null space (set of $|v\rangle$ such that $G |v\rangle = 0$). Then the *Positive Inverse* of G is defined as

$$G^\dashv := \Pi \left(G + \Pi^\perp\right)^{-1} \Pi.$$

Equivalently, one could use the spectral decomposition. Given $G = \sum_{i=1}^m \lambda_i |i\rangle \langle i|$ where all $\lambda_i > 0$ without loss of generality,

$$G^\dashv := \sum_{i=1}^m \lambda_i^{-1} |i\rangle \langle i| .$$

The curvature of the ellipsoid at a given point is given by the so-called Weingarten Map (we saw this in Chapter 6). In practice, it is easier to evaluate the Reverse Weingarten Map which is denoted by W and its positive inverse, $W^\dashv$, yields the Weingarten Map. Suppose the ellipsoid corresponds to G. Then, if G and $G^\dashv$ are known then one can find analytic expressions for both W and $W^\dashv$. We defer these derivations to Section D.1, of the Appendix and instead simply state these results collectively.

This is useful because, intuitively, we start with full rank diagonal matrices, $G^{\bar{n}} = X_g$ (and analogously $H^{\bar{n}} = X_h$) which are readily invertible, keeping track of both $W =: G^{\overline{n-1}}$ and $W^\dashv$ lets us apply the same procedure on $G^{\overline{n-1}}$ (we shortly see why the rank drops).

Definition 147 (Normal Function, Reverse Weingarten Map, Inverse of the Weingarten Map, Orthogonal Component)**.** Given a matrix $G \geq 0$, its positive inverse $G^\dashv$ and a vector $|v\rangle$ (such that $G |v\rangle \neq 0$) we define the following functions. We use $\left\langle G^j \right\rangle := \langle v| G^j |v\rangle$.

- The normal function, from $G, |v\rangle$ to a vector $|u\rangle$ is defined as

$$|u(G,|v\rangle)\rangle := \frac{G|v\rangle}{\langle G^2\rangle}.$$

- The Weingarten Map from $G, |v\rangle$ to a matrix $W^{\dashv}$ is defined as

$$W^{\dashv}(G,|v\rangle) := \sqrt{\frac{\langle G\rangle}{\langle G^2\rangle}}\left(G + \frac{\langle G^3\rangle}{\langle G^2\rangle^2} G|v\rangle\langle v|G - \frac{1}{\langle G^2\rangle}\left(G|v\rangle\langle v|G^2 + G^2|v\rangle\langle v|G\right)\right).$$

- The Reverse Weingarten Map from $G, G^{\dashv}, |v\rangle$ to W is defined to be

$$W(G,G^{\dashv},|v\rangle) := \sqrt{\frac{\langle G^2\rangle}{\langle G\rangle}}\left(G^{\dashv} - \frac{|v\rangle\langle v|}{\langle G\rangle}\right).$$

- The Orthogonal Component from $G, |v\rangle$ to $|e\rangle$ is defined to be

$$|e(G,|v\rangle)\rangle := \mathcal{N}\left[|v\rangle - \langle u|v\rangle|u\rangle\right],$$

where $|u\rangle = |u(G,|v\rangle)\rangle$.
The Orthogonal Component from $|v'\rangle, |v\rangle$ to $|e\rangle$ is defined to be

$$\left|e(|v'\rangle, |v\rangle)\right\rangle := \mathcal{N}[|v\rangle - \langle v'|v\rangle|v'\rangle].$$

Evaluating the Weingarten map at a given point of a rotated ellipsoid is the same as evaluating it for the unrotated ellipsoid and then rotating it. The following remark makes this precise.

Remark 148. Let $G \geq 0$ be an $n \times n$ rank k matrix and Q be an isometry from the non-trivial k-dimensional subspace of G to an arbitrary k-dimensional subspace. Then

$$W^{\dashv}(QGQ^T, Q|v\rangle) = QW^{\dashv}(G,|v\rangle)Q^T.$$

The following shows why W necessarily has one less rank compared to G.

Remark 149. With reference to Definition 147, let $W = W(G,G^{\dashv},|v\rangle)$, $W^{\dashv} = W^{\dashv}(G,|v\rangle)$ and $|u\rangle = |u(G,|v\rangle)\rangle$. Then $WG|v\rangle = W|u\rangle = 0$ and $W^{\dashv}G|v\rangle = W^{\dashv}|u\rangle = 0$. This may be seen by a direct computation. Alternatively the proofs in Section D.1 (in particular those of Lemma 187 and Lemma 185), should make it evident.

As motivated in Subsection 1.3.2, (and then extensively used in Chapter 6) our interest in the geometry of ellipsoids stems from the following connection with matrix inequalities. These inequalities appear in EBRM transitions (see Corollary 102). Let $H \geq 0$ and $G \geq 0$. One can rewrite a matrix inequality as follows:

$$\begin{aligned} & H - OGO^T \geq 0 && \\ \iff & \langle s|H|s\rangle - \langle s|OGO^T|s\rangle \geq 0 && \forall\ |s\rangle \\ \iff & \langle s|OGO^T|s\rangle \leq 1 && \forall\ \{|s\rangle \,|\, \langle s|H|s\rangle = 1\}. \end{aligned}$$

From Definition 145 one can interpret the last step as stating that along all directions $|s\rangle$, the ellipsoid corresponding to H will be inside the ellipsoid corresponding to OGO^T. If H and G are fixed, then finding the orthogonal matrix O can be seen as rotating the G ellipsoid into an orientation such that the H ellipsoid stays inside.

§ 7.2 f_0 Unitary | Solution to the f_0-assignment

Recall that a valid function is the same as an EBRM function (see Corollary 102). Given a valid function $t = \sum_i p_{h_i} [\![x_{h_i}]\!] - \sum_i p_{g_i} [\![x_{g_i}]\!]$, it is easy to re-write the matrices that appear in the EBRM description into a form which satisfies[2] $H \geq OGO^T$, $O|v\rangle = |w\rangle$, where $|v\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots)$ and $|w\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots)$ while $H = \text{diag}(x_{h_1}, x_{h_2} \dots)$ and $G = \text{diag}(x_{g_1}, x_{g_2} \dots)$.

As we already saw in Chapter 5, it suffices to restrict to monomial assignments (see Definition 80), i.e. assignments of the form

$$t = \sum_{i=1}^{n} \frac{-(-x_i)^k}{\prod_{j \neq i}(x_j - x_i)}$$

for $0 \leq x_1 < x_2 \cdots < x_n$ with $0 \leq k \leq n-2$, to convert Mochon's games into explicit protocols. We construct the solution for the $k = 0$ case (the f_0-assignment) to motivate and introduce the notation we need for the more general solution.

Also recall that for Mochon's f_0-assignment, we have $\langle x^k \rangle = 0$ for all $0 \leq k \leq n-2$ and that $\langle x^{n-1} \rangle > 0$ (see Lemma 81). We may say that the ellipsoids H and OGO^T touch along the vector $|w\rangle$ if $\langle w| H |w\rangle = \langle w| OGO^T |w\rangle = \langle v| G |v\rangle$. Indeed, since $\langle w| H |w\rangle - \langle v| G |v\rangle = \langle x \rangle = 0$ we can conclude the aforesaid. This in turn means that the normal (see Definition 147) along $|v\rangle$ of the G ellipsoid, $|u(G, |v\rangle)\rangle$ must be mapped to the normal along $|w\rangle$ of the H ellipsoid, $|u(H, |w\rangle)\rangle$, i.e. O must have the form $O = |u(H, |w\rangle)\rangle \langle u(G, |v\rangle)| + Q$ where the Q represents the action from the space orthogonal to $|u(G, |v\rangle)\rangle$ onto the space orthogonal to $|u(H, |w\rangle)\rangle$. Further, because $H \geq OGO^T$ we must have (see Definition 147)

$$W(H, H^\dashv, |w\rangle) \geq QW(G, G^\dashv, |v\rangle)Q^T,$$

the curvature of the H ellipsoid at $|w\rangle$, $W(H, H^\dashv, |w\rangle)$, must be greater than that of the OGO^T ellipsoid along $|v\rangle$, i.e. $QW(G, G^\dashv, |v\rangle)Q^T$ (we used the fact that $W(G, G^\dashv, |v\rangle)\,|u(G, |v\rangle)\rangle = 0$). The component of $|v\rangle$ along $|u(G, |v\rangle)\rangle$ is mapped to $|u(H, |w\rangle)\rangle$ under the action of O (which has so far, only been partially specified). The remaining component is $|e(G, |v\rangle)\rangle$ and analogously for $|w\rangle$, the remaining component is $|e(H, |w\rangle)\rangle$. Using these Ws and $|e\rangle$s as the new matrices and vectors, it turns out that one can apply this argument repeatedly (when the number of points, n, is even) to completely specify O. However, clearly, this notation will rapidly become complicated. We therefore introduce the notion of a *matrix instance* which is similar to the one used in Chapter 6 but instead uses isometries and a slightly different nomenclature.

Definition 150 (Matrix Instance and its properties)**.** Let

- $n \geq k$ be positive integers,
- $\mathcal{H}^{\bar{k}}$ and $\mathcal{G}^{\bar{k}}$ be two k dimensional Hilbert spaces,
- $H \geq 0$, $G \geq 0$ be $n \times n$ non-zero matrices of rank at most k, such that H has support only on $\mathcal{H}^{\bar{k}}$ and analogously G has support only on $\mathcal{G}^{\bar{k}}$,
- $|w\rangle \in \mathcal{H}^{\bar{k}}$ and $|v\rangle \in \mathcal{G}^{\bar{k}}$ be vectors of equal norm, $|u_h\rangle \in \mathcal{H}^{\bar{k}}$ and $|u_g\rangle \in \mathcal{G}^{\bar{k}}$ be vectors with unit norm,

A *matrix instance* is defined to be the tuple $\underline{\mathrm{X}}^{\bar{k}} := (H, G, |w\rangle, |v\rangle)$ and the set of all matrix instances (of $n \times n$ dimensions) is denoted by $\mathbb{X}^n$.

We define the following properties of a matrix instance.

[2] (see the discussion after Theorem 75; we suppressed the details about the dimensions and the spectra of matrices)

- Let $Q : \mathcal{G}^{\bar{k}} \to \mathcal{H}^{\bar{k}}$ be an isometry, i.e. $Q^T Q = \mathbb{I}_h$ and $QQ^T = \mathbb{I}_g$ where $\mathbb{I}_h$ is the identity in $\mathcal{H}^{\bar{k}}$ and similarly $\mathbb{I}_g$ is the identity in $\mathcal{G}^{\bar{k}}$. We say that Q *solves* the *matrix instance* $\underline{\mathrm{X}}^{\bar{k}}$ if and only if

$$\begin{aligned} H &\geq QGQ^T, \\ Q\left|v\right\rangle &= \left|w\right\rangle. \end{aligned}$$

- We say that $\underline{\mathrm{X}}^{\bar{k}}$ satisfies the *contact condition* if and only if $\langle w| H |w\rangle = \langle v| G |v\rangle$.
- We say that $\underline{\mathrm{X}}^{\bar{k}}$ satisfies the *component condition* if and only if $\langle w| H^2 |w\rangle = \langle v| G^2 |v\rangle$.

The definition should appear intuitive, given the preceding discussion. We did not motivate the component condition but this should get clarified shortly. Instead, let us familiarise ourselves with the notation by applying it to the f_0-example. We started with the matrix instance, $\underline{\mathrm{X}}^{\bar{n}} =: (H, G, |w\rangle, |v\rangle)$ and argued that the inner ellipsoid should be more curved than the outer to obtain another matrix instance, with one less dimension. We also evaluated the relevant vectors. We formalise this into what we call a *Weingarten Iteration Map*.

Definition 151 (Weingarten Iteration Map)**.** Consider a matrix instance $\underline{\mathrm{X}}^{\bar{k}} =: \left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right)$ and let (see Definition 147)

$$\begin{aligned} \left|v^{\overline{k-1}}\right\rangle &:= \left|e\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\right\rangle, & \left|w^{\overline{k-1}}\right\rangle &:= \left|e\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right)\right\rangle, \\ G^{\overline{k-1}} &:= W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right), & H^{\overline{k-1}} &:= W^{\dashv}\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right). \end{aligned}$$

Then we define the *Weingarten Iteration Map* $\mathscr{W} : \mathbb{X}^n \to \mathbb{X}^n$ by its action

$$\underline{\mathrm{X}}^{\bar{k}} \mapsto \left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle\right) =: \underline{\mathrm{X}}^{\overline{k-1}}.$$

In our example, note that we relied on the properties of the f_0-assignment, only for establishing that the *contact condition* holds, i.e. $\langle w| H |w\rangle = \langle v| G |v\rangle$. The rest of the argument was actually quite general. We state it in terms of matrix instances and give a more formal proof.

A remark about the notation—while we use the word resolves in both the statement and the proof, we define this properly later; it essentially means that instead of $H \geq QGQ^T$ if we have $H \leq QGQ^T$ then also the argument goes through analogously and it becomes relevant when $k > 0$. We are getting ahead of ourselves as we have not even solved the f_0 assignment which has $k = 0$. We remedy this next.

Lemma 152. *Consider a matrix instance $\underline{\mathrm{X}}^{\bar{k}} := (H, G, |w\rangle, |v\rangle)$ which satisfies both the contact and the component condition. Let $\left|u_h^{\bar{k}}\right\rangle := |u(H, |w\rangle)\rangle$, $\left|u_g^{\bar{k}}\right\rangle := |u(G, |v\rangle)\rangle$ and $\underline{\mathrm{X}}^{\overline{k-1}} := \mathscr{W}(\underline{\mathrm{X}}^{\bar{k}})$ (see Definition 157). We assert that if $Q^{\bar{k}}$ (re)solves the matrix instance $\underline{\mathrm{X}}^{\bar{k}}$ then*

$$Q^{\bar{k}} = \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right| + Q^{\overline{k-1}}, \tag{7.2}$$

where $Q^{\overline{k-1}}$ (re)solves the matrix instance $\underline{\mathrm{X}}^{\overline{k-1}}$.

Proof. Let $\left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right) := \underline{\mathrm{X}}^{\bar{k}}$ and $\left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle\right) := \underline{\mathrm{X}}^{\overline{k-1}}$. Using the ellipsoid picture (see Section 7.1) for the matrix inequality $H^{\bar{k}} \geq Q^{\bar{k}} G^{\bar{k}} \left(Q^{\bar{k}}\right)^T$ it is clear that the

ellipsoid corresponding to $H^{\bar{k}}$ is contained inside the ellipsoid corresponding to $Q^{\bar{k}}G^{\bar{k}}\left(Q^{\bar{k}}\right)^T$. The two ellipsoids touch along the $\left|w^{\bar{k}}\right\rangle$ direction if and only if

$$\left\langle w^{\bar{k}}\right| H^{\bar{k}} \left|w^{\bar{k}}\right\rangle = \left\langle w^{\bar{k}}\right| Q^{\bar{k}}G^{\bar{k}}\left(Q^{\bar{k}}\right)^T \left|w^{\bar{k}}\right\rangle = \left\langle v^{\bar{k}}\right| G^{\bar{k}} \left|v^{\bar{k}}\right\rangle$$

(the last step follows from noting $Q^{\bar{k}}\left|v^{\bar{k}}\right\rangle = \left|w^{\bar{k}}\right\rangle$ and the fact that $Q^{\bar{k}}$ is an isometry). This is precisely the contact condition (which is given to hold). The component condition ensures that the components of the probability vectors along their respective normals are the same, viz. $\left\langle w^{\bar{k}}|u_h^{\bar{k}}\right\rangle = \left\langle v^{\bar{k}}|u_g^{\bar{k}}\right\rangle$ (see Lemma 184). From this we can deduce the following three necessary conditions.

First, that Equation (7.2) holds. Indeed, the normal along $\left|w^{\bar{k}}\right\rangle$ (see Lemma 184) of the ellipsoid $H^{\bar{k}}$ and that of the ellipsoid $Q^{\bar{k}}G^{\bar{k}}Q^{\bar{k}T}$ must be the same. This in turn means that $Q^{\bar{k}}$ must map the normal $\left|u_g^{\bar{k}}\right\rangle$ along $\left|v^{\bar{k}}\right\rangle$ of the ellipsoid $G^{\bar{k}}$ to the normal $\left|u_h^{\bar{k}}\right\rangle$ along $|w\rangle$ of the ellipsoid $H^{\bar{k}}$, viz. $\left|u_g^{\bar{k}}\right\rangle := \left|u\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\right\rangle \mapsto \left|u_h^{\bar{k}}\right\rangle := \left|u\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right)\right\rangle$ (see Definition 147). Consequently,

$$Q^{\bar{k}} = \left|u_h^{\bar{k}}\right\rangle\left\langle u_g^{\bar{k}}\right| + Q^{\overline{k-1}}, \tag{7.3}$$

where $Q^{\overline{k-1}} : \mathcal{G}^{\overline{k-1}} \to \mathcal{H}^{\overline{k-1}}$ is an isometry as the action on the normals is completely determined.

Second, note that the curvature along $\left|w^{\bar{k}}\right\rangle$ of the ellipsoid $H^{\bar{k}}$ must be greater than that of the ellipsoid $Q^{\bar{k}}G^{\bar{k}}\left(Q^{\bar{k}}\right)^T$ (along the same direction), viz.

$$\begin{aligned}
H^{\overline{k-1}} = W^{\dashv}\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right) &\geq W^{\dashv}\left(Q^{\bar{k}}G^{\bar{k}}\left(Q^{\bar{k}}\right)^T, Q^{\bar{k}}\left|v^{\bar{k}}\right\rangle\right) \\
&= Q^{\bar{k}}W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\left(Q^{\bar{k}}\right)^T \\
&= Q^{\overline{k-1}}\underbrace{W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)}_{=G^{\overline{k-1}}}\left(Q^{\overline{k-1}}\right)^T \quad \because \quad W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\left|u_g^{\bar{k}}\right\rangle = 0; \\
&= Q^{\overline{k-1}}G^{\overline{k-1}}\left(Q^{\overline{k-1}}\right)^T. \qquad \text{see 149}
\end{aligned}$$

Finally, since $Q^{\bar{k}}\left|v^{\bar{k}}\right\rangle = \left|w^{\bar{k}}\right\rangle$ by multiplying a projector on both sides, it follows that $\left(\mathbb{I}_h^{\bar{k}} - \left|u_h^{\bar{k}}\right\rangle\left\langle u_h^{\bar{k}}\right|\right)Q^{\bar{k}}\left|v^{\bar{k}}\right\rangle = \left(\mathbb{I}_h^{\bar{k}} - \left|u_h^{\bar{k}}\right\rangle\left\langle u_h^{\bar{k}}\right|\right)\left|w^{\bar{k}}\right\rangle$. Using

$$\left(\mathbb{I}_h^{\bar{k}} - \left|u_h^{\bar{k}}\right\rangle\left\langle u_h^{\bar{k}}\right|\right)Q^{\bar{k}} = \left(\mathbb{I}_h^{\bar{k}} - \left|u_h^{\bar{k}}\right\rangle\left\langle u_h^{\bar{k}}\right|\right)Q^{\bar{k}}\left(\mathbb{I}_g^{\bar{k}} - \left|u_g^{\bar{k}}\right\rangle\left\langle u_g^{\bar{k}}\right|\right)$$

in the LHS (follows from Equation (7.3)) and Definition 147 for $|e(.,.)\rangle$, one obtains the equation $Q^{\overline{k-1}}\left|v^{\overline{k-1}}\right\rangle = \left|w^{\overline{k-1}}\right\rangle$. These show that $Q^{\overline{k-1}}$ indeed solves $\underline{X}^{\overline{k-1}}$. Changing the direction of the inequality doesn't change any other argument, which also proves the resolve case (introduced later). □

7.2.1 The Balanced Case

Proposition 153 (The balanced f_0 Solution). *Let $t = h - g = \sum_{i=1}^{2n} p_i [\![x_i]\!]$ be Mochon's f_0 assignment for the set of real numbers $0 \leq x_1 < x_2 \cdots < x_{2n}$. Let $h = \sum_{i=1}^{n} p_{h_i} [\![x_{h_i}]\!]$, $g = \sum_{i=1}^{n} p_{g_i} [\![x_{g_i}]\!]$*

where p_{h_i} and p_{g_i} are strictly positive, and $\{x_{h_i}\}$ and $\{x_{g_i}\}$ are all distinct. Consider the matrix instance $\underline{X}$:=$(X_h, X_g, |w\rangle, |v\rangle)$ where $X_h \doteq diag(x_{h_1}, x_{h_2} \dots x_{h_n})$, $X_g \doteq diag(x_{g_1}, x_{g_2} \dots x_{g_n})$, $|w\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \dots \sqrt{p_{h_n}})^T$, $|v\rangle \doteq \left(\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \dots \sqrt{p_{g_n}}\right)^T$. The orthogonal matrix

$$O = \sum_{k=1}^{n} \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right|$$

solves $\underline{X} =: \underline{X}^{\bar{n}}$ (see Definition 150) where the Weingarten Iteration Map (see Definition 157) is used to evaluate $\underline{X}^{\overline{k-1}} = \mathscr{W}(\underline{X}^{\bar{k}})$ which in turn is used to obtain $\left|u_h^{\bar{k}}\right\rangle = \left|(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle)\right\rangle$ and $\left|u_g^{\bar{k}}\right\rangle = \left|(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle)\right\rangle$ for all k (where $\left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right) := \underline{X}^{\bar{k}}$), starting from $k = n$.

To prove Proposition 153, we use the following lemma which follows from Lemma 193 and Lemma 196 (proved in Section D.5 of the Appendix).

Lemma 154 (Up Contact/Component Lemma)**.** *Consider the matrix instance $\underline{X}^{\bar{n}} := (H^{\bar{n}}, G^{\bar{n}}, \left|w^{\bar{n}}\right\rangle, \left|v^{\bar{n}}\right\rangle)$. Suppose the Weingarten Iteration Map (see Definition 157) is applied l times to obtain*

$$\underline{X}^{\overline{n-l}} := \left(H^{\overline{n-l}}, G^{\overline{n-l}}, \left|w^{\overline{n-l}}\right\rangle, \left|v^{\overline{n-l}}\right\rangle\right).$$

Then,

$$\left\langle v^{\overline{n-l}}\right| \left(G^{\overline{n-l}}\right)^m \left|v^{\overline{n-l}}\right\rangle = r\left(\left\langle (G^{\bar{n}})^{m-1}\right\rangle, \left\langle (G^{\bar{n}})^{m}\right\rangle \dots, \left\langle (G^{\bar{n}})^{2l+m}\right\rangle\right),$$

where $m \geq 1$ and r is a multi-variate function which does not have an implicit dependence on $\left\langle (G^{\bar{n}})^i\right\rangle := \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle$ for any i. The corresponding statement involving H's and $|w\rangle$'s also holds, i.e.

$$\left\langle w^{\overline{n-l}}\right| \left(H^{\overline{n-l}}\right)^m \left|w^{\overline{n-l}}\right\rangle = r\left(\left\langle (H^{\bar{n}})^{m-1}\right\rangle, \left\langle (H^{\bar{n}})^{m}\right\rangle \dots, \left\langle (H^{\bar{n}})^{2l+m}\right\rangle\right).$$

This lemma relates the contact condition of the lth matrix instance, i.e. the matrix instance obtained after applying the Weingarten Iteration Map l times, to the expectation values associated with the first matrix instance. These expectation values, for the f_0 solution, are $\left\langle x^k\right\rangle = \left\langle (H^{\bar{n}})^k\right\rangle - \left\langle (G^{\bar{n}})^k\right\rangle = 0$ for all $0 \leq k \leq n-2$. This in turn means (details below) that the contact condition also holds for the lth matrix instance, thereby allowing one to repeatedly use Lemma 152 to determine the solution, O.

Proof of Proposition 153. We have already done most of the work (by proving Lemma 152 and Lemma 154). Now only a counting argument remains. At the base level, we have the matrix instance $\underline{X} =: \underline{X}^{\bar{n}} =: (H^{\bar{n}}, G^{\bar{n}}, \left|w^{\bar{n}}\right\rangle, \left|v^{\bar{n}}\right\rangle)$. To use the Weingarten iteration once, we must show that $\underline{X}^{\bar{n}}$ satisfies the contact condition (see Definition 155 and Lemma 152), viz.

$$\left\langle w^{\bar{n}}\right| H^{\bar{n}} \left|w^{\bar{n}}\right\rangle - \left\langle v^{\bar{n}}\right| G^{\bar{n}} \left|v^{\bar{n}}\right\rangle = \left\langle H^{\bar{n}}\right\rangle - \left\langle G^{\bar{n}}\right\rangle = \sum_{i=1}^{n} p_{h_i} x_{h_i} - \sum_{i=1}^{n} p_{g_i} x_{g_i} = \sum_{i=1}^{n} p_i x_i = \langle x\rangle$$

vanishes which it does due to Lemma 81. After iterating for l steps, suppose the matrix instance one obtains is $\underline{X}^{\overline{n-l}}$. To check if another Weingarten iteration is possible, we must check if the contact condition holds, i.e. if

$$\left\langle w^{\overline{n-l}}\right| H^{\overline{n-l}} \left|w^{\overline{n-l}}\right\rangle - \left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle =$$
$$r\left(\left\langle (H^{\bar{n}})^1\right\rangle, \left\langle (H^{\bar{n}})^2\right\rangle \dots, \left\langle (H^{\bar{n}})^{2l+1}\right\rangle\right) - r\left(\left\langle (G^{\bar{n}})^1\right\rangle, \left\langle (G^{\bar{n}})^2\right\rangle \dots, \left\langle (G^{\bar{n}})^{2l+1}\right\rangle\right)$$

vanishes. We used Lemma 154 (with $m = 1$) to obtain the RHS. Note that

$$\left\langle (H^{\bar{n}})^k \right\rangle - \left\langle (G^{\bar{n}})^k \right\rangle = \left\langle x^k \right\rangle . \tag{7.4}$$

If $2l + 1 \leq 2n - 2$ then from Lemma 81 it follows that both terms become identical and hence the difference indeed vanishes.[3] A similar argument can be used to obtain the condition $2l + 2 \leq 2n - 2$ which corresponds to the component condition (see Definition 155). Assuming $O =: O^{\bar{n}}$ solves $\underline{\mathrm{X}}^{\bar{n}}$, until $l = n - 2$ (included), one can iterate (using the Weingarten Iteration Map, $\mathscr{W}$, and the Normal Initialisation Map, $\mathscr{U}$) to obtain $\left|u_h^{\bar{n}}\right\rangle, \left|u_h^{\overline{n-1}}\right\rangle, \ldots, \left|u_h^{\overline{n-l}}\right\rangle, \ldots, \left|u_h^{\bar{1}}\right\rangle$ and similarly $\left|u_g^{\bar{n}}\right\rangle, \left|u_g^{\overline{n-1}}\right\rangle, \ldots, \left|u_g^{\overline{n-l}}\right\rangle, \ldots, \left|u_g^{\bar{1}}\right\rangle$ which completely determine $O^{\bar{n}}$.

It only remains to prove that there exists an O which solves the matrix instance $\underline{\mathrm{X}}^{\bar{n}}$. We outline the proof in Subsection D.4 in the Appendix. A more formal argument is given when we discuss the EMA algorithm (see Subsection 6.3.3, Chapter 6). □

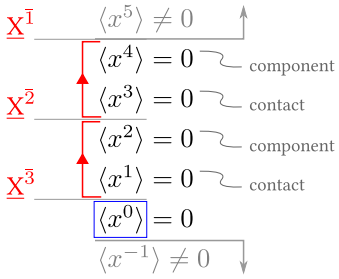


Figure 7.1: Power diagram for a balanced f_0 assignment with $2n = 6$ points. Starting upwards from $\langle x^0 \rangle$, two iterations are completed before encountering the instance where the contact condition does not hold and the normals do not match.

It is helpful to represent the main argument succinctly through a diagram (see Figure 7.1). We start right above $\langle x^0 \rangle$ with the matrix instance $\underline{\mathrm{X}}^{\bar{n}}$. Set $n = 3$ for concreteness. The contact condition at this step corresponds to $\langle x^1 \rangle = 0$, which is true as the power is less than or equal to $2n - 2$ (here $2n - 2 = 4$; see Lemma 81). We can thus apply the Weingarten iteration (see Definition 157) and this is indicated by the arrow[4] from $\langle x^1 \rangle$ to $\langle x^2 \rangle$. This yields $\underline{\mathrm{X}}^{\overline{n-1}}$ and we can proceed with checking if $\langle x^3 \rangle = 0$, which is true as the power is ≤ 4, and therefore we can again iterate to obtain $\underline{\mathrm{X}}^{\overline{n-2}}$, which in this illustration is $\underline{\mathrm{X}}^{\bar{1}}$. At this point, we have solved the problem as we can evaluate $\left|u_h^{\bar{3}}\right\rangle, \left|u_h^{\bar{2}}\right\rangle, \left|u_h^{\bar{1}}\right\rangle$ and $\left|u_g^{\bar{3}}\right\rangle, \left|u_g^{\bar{2}}\right\rangle, \left|u_g^{\bar{1}}\right\rangle$ form $\underline{\mathrm{X}}^{\bar{3}}, \underline{\mathrm{X}}^{\bar{2}}, \underline{\mathrm{X}}^{\bar{1}}$ respectively to write $O = \sum_{k=1}^{3} \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right|$. Note that having an even number of total points, $x_1 < x_2 \cdots < x_{2n}$, ensures that there is a proper *alignment* in the diagram, e.g. the contact condition for $\underline{\mathrm{X}}^{\bar{2}}$ corresponds to $\langle x^3 \rangle = 0$ and the component condition corresponds to $\langle x^4 \rangle = 0$ both of which hold. As we saw in the proof, the contact condition essentially requires that the component of $\left|w^{\bar{k}}\right\rangle$ along $\left|u_h^{\bar{k}}\right\rangle$ is the same as the component of $\left|v^{\bar{k}}\right\rangle$ along $\left|u_g^{\bar{k}}\right\rangle$. If this does not hold, then we would no longer have $O\left|v\right\rangle = \left|w\right\rangle$ which not only means we don't have a solution, it also means that our reasoning, which relied on this assumption, leads to false conclusion.

This is relevant because when the number of points is odd the component condition ceases to hold at the last step (while the contact condition continues to hold). This in turn means that one can no longer apply the Weingarten Iteration Map as the premise for Lemma 152 is not true. We have already encountered this situation in Chapter 6 and the wiggle-v method we used then also works in this case.

[3] The number of points here is $2n$; in the Lemma they are denoted by n.

[4] It, strictly speaking, goes from below $\langle x^1 \rangle$ to above $\langle x^2 \rangle$; the idea was just to indicate the inclusion of the two terms for the matrix instance $\underline{\mathrm{X}}^{\bar{3}}$.

Nevertheless, let us try to recall what the situation corresponds to geometrically. So far, we reasoned that if one ellipsoid is contained inside anther, $H \geq QGQ^T$ and if they touch along a vector, $|w\rangle$, then we can be certain that the normal $|u_g\rangle$ of the G ellipsoid along $|v\rangle = Q^T |w\rangle$ must be mapped to the normal $|u_h\rangle$ of the H ellipsoid along $|w\rangle$ by the isometry Q. This analysis requires that the normal is well defined (which it is if the matrices have finite spectra). However, as we pointed out in Chapter 3 (when we discussed that the cone or set of EBM transitions is not closed) that some valid functions can not be expressed by matrices (EBM) with a finite spectrum and merge was an example. Geometrically, imagine the situation where the QGQ^T ellipsoid is a circle and the H ellipsoid is a line and the $|w\rangle$ vector points along this line (see Figure 7.2; image on the right). The normal to the H ellipsoid along the point of contact, i.e. normal to the line, can have an arbitrary component along the vector perpendicular to $|w\rangle$. If the line is seen as an approximation to a squeezed circle, then it is clear that a very small *wiggle* in $|w\rangle$ can significantly affect the normal. As we have already seen this more precisely in Chapter 6 (and we discuss this again in the proofs to follow), we content ourselves with the observation that there is a freedom in the choice of the normal. We can fix this freedom by requiring that the component condition is satisfied. Denote the direction of infinite curvature (in our "circle-line" example it was the vector perpendicular to $|w\rangle$), by $|t_h\rangle$. The freedom in correcting the normal can be expressed by parametrising it as

$$\left|u_h'\right\rangle := \cos\theta \left|u_h\right\rangle + \sin\theta \left|t_h\right\rangle \tag{7.5}$$

where $|u_h\rangle := |u(H, |w\rangle)\rangle$. Enforcing the component condition, i.e. $\langle w|u_h\rangle = \langle v|u_g\rangle$, fixes θ, the parameter which completely specifies the corrected normal, $|u_h'\rangle$. One can now apply Lemma 152 by using $|u_h'\rangle$ instead of $|u_h\rangle$. Informally then, this also demonstrates how one can construct the solution to the f_0-assignment when the number of points is odd. However, this is as far as we can go without formalising the preceding results.

§ 7.3 Extended Matrix Instances

For the Weingarten Iteration Map, we saw that knowing the matrix instance is sufficient. However, as we saw above, when one of the ellipsoids had infinite curvature, one needed to evaluate the corresponding normal differently. To this end, we define an *Extended Matrix Instance*. The idea is to construct an object which holds certain quantities derived from the initial matrix instance, i.e. normals and inverses of the matrices (specifying the matrix instance[5]). The exact formulae for these derived quantities are defined as maps which act on these objects and populate the entries corresponding to the derived quantities.

Definition 155 ((Extended) Matrix Instance and its properties)**.** Let

- $n \geq k$ be positive integers,
- $\mathcal{H}^{\bar{k}}$ and $\mathcal{G}^{\bar{k}}$ be two k dimensional Hilbert spaces,
- S_h be the set of $n \times n$ non-zero matrices of rank at most k with support only on $\mathcal{H}^{\bar{k}}$, i.e.

$$S_h := \{n \times n \text{ matrices } M : M \text{ has rank at most } k, M \geq 0 \text{ and } M \text{ has support only on } \mathcal{H}^{\bar{k}}\}$$

and analogously,

$$S_g := \{n \times n \text{ matrices } M : M \text{ has rank at most } k, M \geq 0 \text{ and } M \text{ has support only on } \mathcal{G}^{\bar{k}}\}$$

[5] This we do from hindsight but can be easily motivated—for a monomial assignment (for a monomial of order $k > 0$), we start higher up in the power diagram (contrast Figure 7.1 and Figure 7.4), and to go lower we need matrix inverses.

- $H \in S_h$, $G \in S_g$, $H_{\text{inv}} \in S_h \cup \{[.]\}$, $G_{\text{inv}} \in S_g \cup \{[.]\}$
- $|w\rangle \in \mathcal{H}^{\bar{k}}$ and $|v\rangle \in \mathcal{G}^{\bar{k}}$ be vectors of equal norm,

$$|u_h\rangle \in \{|u\rangle \in \mathcal{H}^{\bar{k}} : \langle u|u\rangle = 1\} \cup \{|.\rangle\}$$

and

$$|u_g\rangle \in \{|u\rangle \in \mathcal{G}^{\bar{k}} : \langle u|u\rangle = 1\} \cup \{|.\rangle\},$$

A *matrix instance* is defined to be the tuple $\underline{\mathrm{X}}^{\bar{k}} := (H, G, |w\rangle, |v\rangle)$ while an *extended matrix instance* is defined to be the tuple $\underline{\mathrm{M}}^{\bar{k}} := \underline{\mathrm{X}}^{\bar{k}} \oplus (H_{\text{inv}}, G_{\text{inv}}, |u_h\rangle, |u_g\rangle)$.

The extended matrix instance is *partially specified* if H_{inv} or G_{inv} equal $[.]$ or if $|u_h\rangle$ or $|u_g\rangle$ equal $|.\rangle$. We say that an extended matrix instance is *completely specified* if it is not partially specified.

The set of all matrix instances (of $n \times n$ dimensions) is denoted by $\mathbb{X}^n$ and that of extended matrix instances is denoted by $\mathbb{M}^n$. We now define some properties of the (extended) matrix instance.

- Let $Q : \mathcal{G}^{\bar{k}} \to \mathcal{H}^{\bar{k}}$ be an isometry, i.e. $Q^T Q = \mathbb{I}_h$ and $QQ^T = \mathbb{I}_g$ where $\mathbb{I}_h$ is the identity in $\mathcal{H}^{\bar{k}}$ and similarly $\mathbb{I}_g$ is the identity in $\mathcal{G}^{\bar{k}}$. We say that Q *solves* the *matrix instance* $\underline{\mathrm{X}}^{\bar{k}}$ if and only if

$$\begin{aligned} H &\geq QGQ^T, \\ Q\,|v\rangle &= |w\rangle\,. \end{aligned}$$

  Similarly we say that Q *resolves* (reverse solves) the matrix instance if and only if

$$\begin{aligned} H &\leq QGQ^T, \\ Q\,|v\rangle &= |w\rangle\,. \end{aligned}$$

- We say that $\underline{\mathrm{X}}^{\bar{k}}$ satisfies the *contact condition* if and only if $\langle w|\, H\, |w\rangle = \langle v|\, G\, |v\rangle$. Similarly for $\underline{\mathrm{M}}^{\bar{k}}$.
- We say that $\underline{\mathrm{X}}^{\bar{k}}$ satisfies the *component condition* if and only if $\langle w|\, H^2\, |w\rangle = \langle v|\, G^2\, |v\rangle$. Similarly for $\underline{\mathrm{M}}^{\bar{k}}$.
- We say that $\underline{\mathrm{X}}^{\bar{k}}$ has *wiggle-w room* (ϵ) *along* $|t_h\rangle$ if and only if H has an eigenvector $|t_h\rangle$ with eigenvalue $1/\epsilon$ which has no overlap with $|w\rangle$, viz. $H\,|t_h\rangle = \epsilon^{-1}\,|t_h\rangle$ and $\langle w|t_h\rangle = 0$. Similarly, we say that $\underline{\mathrm{X}}^{\bar{k}}$ has *wiggle-v room* (ϵ) along $|t_g\rangle$ if and only if G has an eigenvector $|t_g\rangle$ with eigenvalue $1/\epsilon$ which has no overlap with $|v\rangle$, viz. $G\,|t_g\rangle = \epsilon^{-1}\,|t_g\rangle$ and $\langle v|t_g\rangle = 0$. For brevity, we say $\underline{\mathrm{X}}^{\bar{k}}$ has *wiggle-w/v room.*

We included the abstract symbols $[.]$ and $|.\rangle$ to allow the matrices and vectors to be specified separately. We also repeated the definition of a matrix instance for ease of reference. We used H_{inv} and G_{inv} to indicate that we extend a matrix instance with $H_{\text{inv}} = H^{\dashv}$ and $G_{\text{inv}} = G^{\dashv}$. We did not define them this way right away because we want H_{inv} and G_{inv} to be explicit functions of H and G respectively, i.e. $H_{\text{inv}} = H^{\dashv}(H)$ and $G_{\text{inv}} = G^{\dashv}(G)$. We specify them using maps, discussed next. We introduced the notion of wiggle-v/w room with ϵs to formalise the limiting argument in the proof later.

We formalise the procedure we followed for the balanced f_0 solution (i.e. the one with even points; the easy case). We hope this exercise clarifies the notation, even though we are being slightly redundant.

§ 7.4 Weingarten Iteration | Isometric Iteration using the Weingarten Map

7.4.1 The Finite Case

We discuss two maps to extend the matrix instance which may together be used to completely specify the extended matrix instance. The first—Normal Initialisation Map—evaluates the normals associated with a(n extended) matrix instance, resulting in a (possibly partially specified) extended matrix instance. The second—Weingarten Initialisation Map—takes a rank k (extended) matrix instance and constructs a rank $k-1$ (extended) matrix instance. Lemma 152 relates the solution of these two matrix instances under certain conditions. These results (and their extensions) are the workhorses of our construction. We successively reduce the problem, while retaining analytic expressions for all the quantities involved, until the problem is solved. The conditions we mentioned can be shown to hold for Mochon's assignments, as we already saw was the case for the balanced f_0-assignment.

We use a single variable to specify a matrix instance, e.g. $\underline{\mathrm{X}}$, instead of specifying the four components which define it explicitly every time. In fact, we use the (extended) matrix instance to define the components which are relevant to the discussion, using the $*$ notation. The following should serve as an example.

Definition 156 (Normal Initialisation Map)**.** Given a matrix instance $\underline{\mathrm{X}}^{\bar{k}} =: (H, G, |w\rangle, |v\rangle)$, $H^{\dashv}$, and $G^{\dashv}$ the *normal initialisation map* $\mathscr{U} : \mathbb{X}^n \to \mathbb{M}^n$ (see Definition 155) is defined by its action

$$\underline{\mathrm{X}}^{\bar{k}} \mapsto \underline{\mathrm{X}}^{\bar{k}} \oplus (H^{\dashv}, G^{\dashv}, |u(H, |w\rangle)\rangle, |u(G, |v\rangle)\rangle).$$

Given an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}}$, let $(*, \cdots *, |u_h\rangle, |u_g\rangle) := \underline{\mathrm{M}}^{\bar{k}}$ (see Equation (7.1)). The *normal initialisation map* $\mathscr{U} : \mathbb{M}^n \to \mathbb{M}^n$ leaves all components of $\underline{\mathrm{M}}^{\bar{k}}$ unchanged, except $|u_h\rangle$ and $|u_g\rangle$ which are mapped as (see Definition 147):

$$\begin{aligned} |u_h\rangle &\mapsto |u(H, |w\rangle)\rangle \\ |u_g\rangle &\mapsto |u(G, |v\rangle)\rangle . \end{aligned}$$

The *Normal Initialisation Map* simply formalises the evaluation of the normal, as we did in the f_0-solution (and combines the inverses into a matrix instance for initialisation).

Definition 157 (Weingarten Iteration Map)**.** Consider a matrix instance $\underline{\mathrm{X}}^{\bar{k}} =: \left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right)$ and let (see Definition 147)

$$\begin{aligned} \left|v^{\overline{k-1}}\right\rangle &:= \left|e\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\right\rangle , & \left|w^{\overline{k-1}}\right\rangle &:= \left|e\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right)\right\rangle , \\ G^{\overline{k-1}} &:= W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right) , & H^{\overline{k-1}} &:= W^{\dashv}\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right) . \end{aligned}$$

Then we define the *Weingarten Iteration Map* $\mathscr{W} : \mathbb{X}^n \to \mathbb{X}^n$ by its action

$$\underline{\mathrm{X}}^{\bar{k}} \mapsto \left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle\right) =: \underline{\mathrm{X}}^{\overline{k-1}}.$$

Consider an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}} =: \underline{\mathrm{X}}^{\bar{k}} \oplus S$ and let $\left((H^{\bar{k}})^{\dashv}, (G^{\bar{k}})^{\dashv}, *, *\right) := S$ (see Equation (7.1)). Let (see Definition 147)

$$(G^{\overline{k-1}})^{\dashv} := W\left(G^{\bar{k}}, (G^{\bar{k}})^{\dashv}, \left|v^{\bar{k}}\right\rangle\right) , \qquad (H^{\overline{k-1}})^{\dashv} := W\left(H^{\bar{k}}, (H^{\bar{k}})^{\dashv}, \left|w^{\bar{k}}\right\rangle\right) .$$

Then we define the *Weingarten Iteration Map* $\mathscr{W} : \mathbb{M}^n \to \mathbb{M}^n$ by its action

$$\underline{\mathrm{M}}^{\bar{k}} \mapsto \underline{\mathrm{X}}^{\overline{k-1}} \oplus \left((H^{\overline{k-1}})^{\dashv}, (G^{\overline{k-1}})^{\dashv}, |.\rangle, |.\rangle \right) =: \underline{\mathrm{M}}^{\overline{k-1}}.$$

We had already defined the *Weingarten Iteration Map* (see Definition 151). We now extended it to handle extended matrix instances. Note that the extended matrix instance one obtains after applying the Weingarten Iteration Map is only partially specified, i.e. it leaves the normal vectors unspecified. For completeness, note that these definitions are justified by Lemma 152.

Recall from our discussion of the unbalanced f_0-solution (the one with odd number points, the harder case where we had to use wiggle-w/v) that how the normal vectors are evaluated depended on whether we were at the last step or not. This notation, by allowing one to leave the normal vectors unspecified, allows one to apply the relevant procedure afterwards.

7.4.2 The Divergent Case

We now formalise the wiggle-w/v part which allows us to handle infinite curvatures which necessarily appear in some cases such as the unbalanced f_0-solution. The Normal Initialisation Map for this case essentially captures Definition 156.

Definition 158 (Wiggle-w/v Normal Initialisation Map)**.** Consider a matrix instance $\underline{\mathrm{X}}^{\bar{k}}$, let $(H, G, |w\rangle, |v\rangle) := \underline{\mathrm{X}}^{\bar{k}}$ with wiggle-w room along $|t_h\rangle$ (see Definition 155). The *Wiggle-w Normal Initialisation Map* $\mathscr{U}_w : \mathbb{X}^n \to \mathbb{M}^n$ is defined by its action

$$\underline{\mathrm{X}}^{\bar{k}} \mapsto \underline{\mathrm{X}}^{\bar{k}} \oplus ([.], [.], \cos\theta\, |u(H, |w\rangle)\rangle + \sin\theta\, |t_h\rangle, |u(G, |v\rangle)\rangle)$$

where $\cos\theta := \langle v|u(G, |v\rangle)\rangle \,/\, \langle w|u(H, |w\rangle)\rangle$ (see Definition 156).

Given an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}}$, let $(*, \cdots *, |u_h\rangle, |u_g\rangle) := \underline{\mathrm{M}}^{\bar{k}}$ (see Equation (7.1)), the *Wiggle-w Normal Initialisation Map* $\mathscr{U}_w : \mathbb{M}^n \to \mathbb{M}^n$ is defined by its action on $|u_h\rangle$ and $|u_g\rangle$ (see Definition 156) as

$$\begin{aligned} |u_h\rangle &\mapsto \cos\theta\, |u(H, |w\rangle)\rangle + \sin\theta\, |t_h\rangle \\ |u_g\rangle &\mapsto |u(G, |v\rangle)\rangle\,. \end{aligned}$$

Similarly, consider a matrix instance $(H, G, |w\rangle, |v\rangle) := \underline{\mathrm{X}}^{\bar{k}}$ with wiggle-v room along $|t_g\rangle$ (see Definition 155). The *Wiggle-v Normal Initialisation Map* $\mathscr{U}_v : \mathbb{X}^n \to \mathbb{M}^n$ is defined by its action

$$\underline{\mathrm{X}}^{\bar{k}} \mapsto \underline{\mathrm{X}}^{\bar{k}} \oplus ([.], [.], |u(H, |w\rangle)\rangle, \cos\theta\, |u(G, |w\rangle)\rangle + \sin\theta\, |t_g\rangle)$$

where $\cos\theta := \langle w|u(H, |w\rangle)\rangle \,/\, \langle v|u(G, |v\rangle)\rangle$ (see Definition 156).

Given an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}}$, let $(*, \cdots *, |u_h\rangle, |u_g\rangle) := \underline{\mathrm{M}}^{\bar{k}}$ (see Equation (7.1)), the *Wiggle-v Normal Initialisation Map* $\mathscr{U}_v : \mathbb{M}^n \to \mathbb{M}^n$ is defined by its action on $|u_h\rangle$ and $|u_g\rangle$ (see Definition 156) as

$$\begin{aligned} |u_h\rangle &\mapsto |u(H, |w\rangle)\rangle \\ |u_g\rangle &\mapsto \cos\theta\, |u(G, |v\rangle)\rangle + \sin\theta\, |t_g\rangle\,. \end{aligned}$$

While the *Wiggle-w/v Normal Initialisation Map* should have appeared to be quite straightforward, the *Wiggle-w/v Iteration Map* deserves some explanation before being defined. Recall that the Weingarten Map (and the Reverse Weingarten Map as well; see Definition 147) took as input a matrix and the position vector along which to evaluate the Weingarten Map. However, we know from the unbalanced f_0-solution that a very small change in the position vector can lead a significant change in the normal and therefore also in the calculation of the curvature. Therefore, instead of using the position, we use the corrected normal vector for evaluating the Weingarten map. Recall (see Definition 147) that the normal vector along the direction $|w\rangle$ for the ellipsoid H, is given by $\mathcal{N}(H\,|w\rangle)$. We can run the argument in reverse. Given a normal vector $|u_h\rangle$, the corresponding position vector (or more precisely the direction from the origin to the point at which the normal is given) is simply $\mathcal{N}(H^{\dashv}\,|u_h\rangle)$. It is this that we use for evaluating the Weingarten Map. We now state the definition.

Definition 159 (Wiggle-w/v Iteration Map)**.** Consider an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}}$ and let

$$\left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle, (H^{\bar{k}})^{\dashv}, (G^{\bar{k}})^{\dashv}, \left|u_h^{\bar{k}}\right\rangle, \left|u_g^{\bar{k}}\right\rangle\right) := \underline{\mathrm{M}}^{\bar{k}}.$$

Further, let (see Definition 156)

$$\begin{aligned}
\left|v^{\overline{k-1}}\right\rangle &= \left|e\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right)\right\rangle, & \left|w^{\overline{k-1}}\right\rangle &= \left|e\left(\left|u_h^{\bar{k}}\right\rangle, \left|w^{\bar{k}}\right\rangle\right)\right\rangle,\\
G^{\overline{k-1}} &= W^{\dashv}\left(G^{\bar{k}}, \left|v^{\bar{k}}\right\rangle\right), & H^{\overline{k-1}} &= W^{\dashv}\left(H^{\bar{k}}, \mathcal{N}\left((H^{\bar{k}})^{\dashv}\left|u_h^{\bar{k}}\right\rangle\right)\right),\\
(G^{\overline{k-1}})^{\dashv} &= W\left(G^{\bar{k}}, (G^{\bar{k}})^{\dashv}, \left|v^{\bar{k}}\right\rangle\right), & (H^{\overline{k-1}})^{\dashv} &= W\left(H^{\bar{k}}, (H^{\bar{k}})^{\dashv}, \mathcal{N}\left((H^{\bar{k}})^{\dashv}\left|u_h^{\bar{k}}\right\rangle\right)\right).
\end{aligned}$$

The *Wiggle-w Iteration Map* $\mathscr{W}_w : \mathbb{M}^n \to \mathbb{M}^n$ is defined by its action

$$\underline{\mathrm{M}}^{\bar{k}} \mapsto \left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle, (H^{\overline{k-1}})^{\dashv}, (G^{\overline{k-1}})^{\dashv}, |.\rangle, |.\rangle\right) =: \mathrm{M}^{\overline{k-1}}.$$

Similarly, consider an extended matrix instance $\underline{\mathrm{M}}^{\bar{k}}$ and let

$$\left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle, (H^{\bar{k}})^{\dashv}, (G^{\bar{k}})^{\dashv}, \left|u_h^{\bar{k}}\right\rangle, \left|u_g^{\bar{k}}\right\rangle\right) := \underline{\mathrm{M}}^{\bar{k}}.$$

Further, let (see Definition 156)

$$\begin{aligned}
\left|v^{\overline{k-1}}\right\rangle &= \left|e\left(\left|u_g^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right)\right\rangle, & \left|w^{\overline{k-1}}\right\rangle &= \left|e\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right)\right\rangle,\\
G^{\overline{k-1}} &= W^{\dashv}\left(G^{\bar{k}}, \mathcal{N}\left((G^{\bar{k}})^{\dashv}\left|u_g^{\bar{k}}\right\rangle\right)\right), & H^{\overline{k-1}} &= W^{\dashv}\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right),\\
(G^{\overline{k-1}})^{\dashv} &= W\left(G^{\bar{k}}, (G^{\bar{k}})^{\dashv}, \mathcal{N}\left((G^{\bar{k}})^{\dashv}\left|u_g^{\bar{k}}\right\rangle\right)\right), & (H^{\overline{k-1}})^{\dashv} &= W\left(H^{\bar{k}}, (H^{\bar{k}})^{\dashv}, \left|w^{\bar{k}}\right\rangle\right).
\end{aligned}$$

The *Wiggle-v Iteration Map* $\mathscr{W}_v : \mathbb{M}^n \to \mathbb{M}^n$ is defined by its action

$$\underline{\mathrm{M}}^{\bar{k}} \mapsto \left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle, (H^{\overline{k-1}})^{\dashv}, (G^{\overline{k-1}})^{\dashv}, |.\rangle, |.\rangle\right) =: \mathrm{M}^{\overline{k-1}}.$$

With the notation in place, we can state the analogue of Lemma 152.

Lemma 160. *Consider an extended matrix instance $\underline{M}^{\bar{k}}$ with wiggle-w room ϵ along $\left|t_h^{\bar{k}}\right\rangle$ (see Definition 155). Assume it is completely specified (see Definition 155), and it satisfies both $\mathscr{U}_w(\underline{M}^{\bar{k}}) = \underline{M}^{\bar{k}}$ (see Definition 158) and the contact condition (see Definition 155). Let $\left(*, \cdots *, \left|u_h^{\bar{k}}\right\rangle, \left|u_g^{\bar{k}}\right\rangle\right) := \underline{M}^{\bar{k}}$ and $\underline{M}^{\overline{k-1}} := \mathscr{W}_w(M^{\bar{k}})$ (see Definition 159). We assert that if $Q^{\bar{k}}$ solves $\underline{M}^{\bar{k}}$ in the limit of $\epsilon \to 0$ then*

$$Q^{\bar{k}} = \left|u_h^{\bar{k}}\right\rangle\left\langle u_g^{\bar{k}}\right| + Q^{\overline{k-1}}, \tag{7.6}$$

where $Q^{\overline{k-1}}$ *solves* $\underline{M}^{\overline{k-1}}$.

Similarly, consider an extended matrix instance $\underline{M}^{\bar{k}}$ *with wiggle-v room* ϵ *along* $\left|t_h^{\bar{k}}\right\rangle$ *(see Definition 155). Assume it is completely specified (see Definition 155), and it satisfies both* $\mathscr{U}_v(\underline{M}^{\bar{k}}) = \underline{M}^{\bar{k}}$ *and the contact condition. Let* $\left(*, \cdots *, \left|u_h^{\bar{k}}\right\rangle, \left|u_g^{\bar{k}}\right\rangle\right) := \underline{M}^{\bar{k}}$ *and* $\underline{M}^{\overline{k-1}} := \mathscr{W}_v(M^{\bar{k}})$. *We assert that if* $Q^{\bar{k}}$ *resolves* $\underline{M}^{\bar{k}}$ *in the limit of* $\epsilon \to 0$ *then*

$$Q^{\bar{k}} = \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right| + Q^{\overline{k-1}},$$

where $Q^{\overline{k-1}}$ *resolves* $\underline{M}^{\overline{k-1}}$.

It must be noted that even though the solution works in the limit of $\epsilon \to 0$, this is not unphysical. As was outlined after Theorem 75, it corresponds to allowing projections in the description of the protocol. We already saw see this in detail in Section 6.1, Chapter 6.

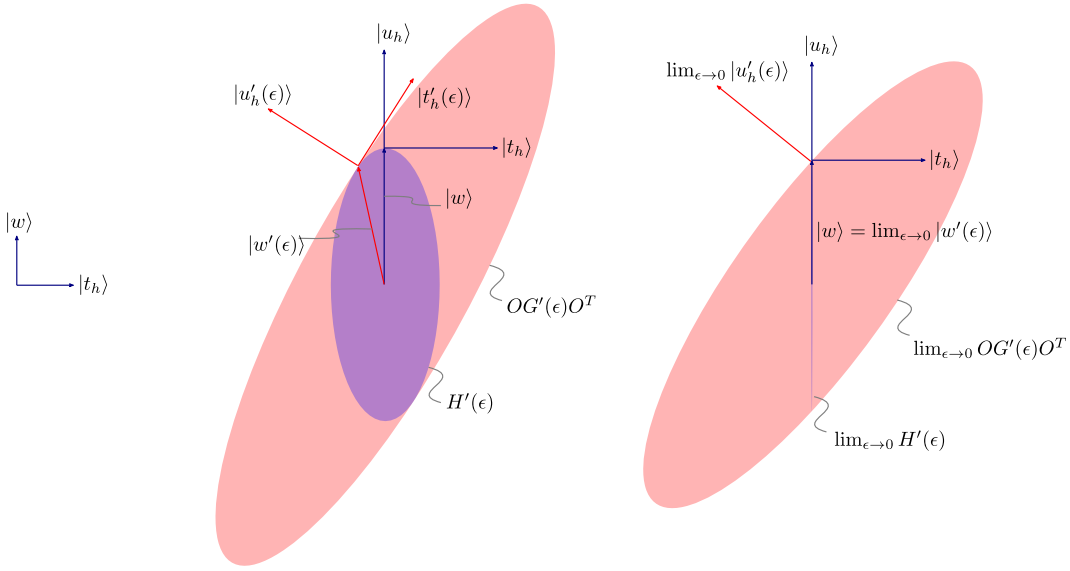


Figure 7.2: The infinite curvature case, where the wiggle-v method is applied. Physically, this translates into using projectors in the protocol.

Proof. We outline the argument in the beginning and then proceed more carefully. The details are slightly distracting and are given later when we discuss the EMA algorithm (see Figure 6.3 and the associated proof of the wiggle-v method in Subsection 6.3.3.2, Chapter 6). The basic idea is that the component of the normal along the $\left|t_h^{\bar{k}}\right\rangle$ direction can be taken to be arbitrary in the limit of $\epsilon \to 0$. To see this, is it helpful to consider a slightly different sequence of matrix instances, parametrised by ϵ, $\underline{X}'^{\bar{k}}(\epsilon) =: \left(H'^{\bar{k}}(\epsilon), G'^{\bar{k}}(\epsilon), \left|w'^{\bar{k}}(\epsilon)\right\rangle, \left|v'^{\bar{k}}(\epsilon)\right\rangle\right)$, which converges as $\epsilon \to 0$ to $\lim_{\epsilon\to 0} \underline{X}^{\bar{k}}(\epsilon) =: \left(H^{\bar{k}}, G^{\bar{k}}, \left|w^{\bar{k}}\right\rangle, \left|v^{\bar{k}}\right\rangle\right)$ (see Figure 7.2). One can use operator monotones (we omit the details here) to construct such a sequence explicitly and show that the solution of all these instances is the same as a function of ϵ. While the parameters specifying the matrix instances converge, the normal $\left|u_h'^{\bar{k}}(\epsilon)\right\rangle = \left|u\left(H'^{\bar{k}}(\epsilon), \left|w'^{\bar{k}}(\epsilon)\right\rangle\right)\right\rangle$, which is a derived quantity, does not converge to $\left|u\left(H^{\bar{k}}, \left|w^{\bar{k}}\right\rangle\right)\right\rangle$, viz.

$$\lim_{\epsilon\to 0}\left|u_h^{\bar{k}}(\epsilon)\right\rangle \neq \left|u\left(H'^{\bar{k}}, \left|w'^{\bar{k}}\right\rangle\right)\right\rangle.$$

This is because a small wiggle in $\left|w^{\bar{k}}\right\rangle$ can significantly affect the calculation of the normal as the curvature along one of the directions diverges. Hence, given $H^{\bar{k}}$ evaluating the normal along $\lim_{\epsilon\to 0}\left|w^{\bar{k}}(\epsilon)\right\rangle$ is not the same as evaluating the normal along $\left|w^{\bar{k}}\right\rangle$.

We can iterate $\underline{\mathrm{X}}'^{\bar{k}}(\epsilon)$ using Definition 157 and Lemma 152, and because the complete solution doesn't depend on ϵ, we can use it to iterate $\underline{\mathrm{X}}^{\bar{k}}$. Since it is along $\left|t_h^{\bar{k}}\right\rangle$ where the curvature diverges as $\epsilon\to 0$, the component of the normal along this direction gets ill-defined. Using the aforesaid reasoning, we can deduce that[6]

$$\lim_{\epsilon\to 0}\left|u_h'^{\bar{k}}(\epsilon)\right\rangle = \cos\theta\left|u\left(H^{\bar{k}},\left|w^{\bar{k}}\right\rangle\right)\right\rangle + \sin\theta\left|t_h^{\bar{k}}\right\rangle,$$

where $\cos\theta$ remains to be determined. The contact condition, $\left\langle u_h'^{\bar{k}}(\epsilon)|w'^{\bar{k}}(\epsilon)\right\rangle = \left\langle u_g'^{\bar{k}}(\epsilon)|v'^{\bar{k}}(\epsilon)\right\rangle$, in the limit $\epsilon\to 0$ becomes

$$\cos\theta\left\langle u\left(H^{\bar{k}},\left|w^{\bar{k}}\right\rangle\right)|w^{\bar{k}}\right\rangle = \left\langle u_g^{\bar{k}}|v^{\bar{k}}\right\rangle$$

(since $\left\langle w^{\bar{k}}|t_h^{\bar{k}}\right\rangle = 0$) thus fixing $\cos\theta$. Defining $\left|u_h^{\bar{k}}\right\rangle := \lim_{\epsilon\to 0}\left|u_h'^{\bar{k}}(\epsilon)\right\rangle$ justifies Definition 158.

Using $\mathscr{W}(\underline{\mathrm{X}}'^{\bar{k}}(\epsilon)) =: \underline{\mathrm{X}}^{\overline{k-1}}(\epsilon) =: \left(H'^{\overline{k-1}}(\epsilon), G'^{\overline{k-1}}(\epsilon), \left|w'^{\overline{k-1}}(\epsilon)\right\rangle, \left|v'^{\overline{k-1}}(\epsilon)\right\rangle\right)$, in the limit $\epsilon\to 0$, we define

$$\underline{\mathrm{X}}^{\overline{k-1}} =: \left(H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle\right).$$

Since the diverging term is in $H'^{\bar{k}}(\epsilon)$ and not in $G'^{\bar{k}}(\epsilon)$ it follows that $G^{\overline{k-1}}$ and $\left|v^{\overline{k-1}}\right\rangle$ can be evaluated using the usual rule specified by the Weingarten Iteration Map, $\mathscr{W}$ on $\underline{\mathrm{X}}^{\bar{k}}$. The relatively non-trivial part is to show that $H^{\overline{k-1}}$ and $\left|w^{\overline{k-1}}\right\rangle$ can be equivalently defined using the correct normal, $\left|u_h^{\bar{k}}\right\rangle$. We use the fact that we can run the following observation backwards: given a direction of contact $|w\rangle$, the normal vector of the ellipsoid represented by H is along $H|w\rangle$, viz. given a normal vector $|u\rangle$, one can obtain the (direction of) point of contact as $H^{-1}|u\rangle$. As we argued above, $\left|w^{\bar{k}}\right\rangle$ can not be reliably used to derive quantities and therefore $\left|u_h^{\bar{k}}\right\rangle$ (together with the said observation) is used to evaluate the Weingarten map[7], as defined in Definition 159.

If Q solves $\underline{\mathrm{X}}'^{\bar{k}}(\epsilon)$ then from Lemma 152 we know that $Q^{\bar{k}} = \left|u_h'^{\bar{k}}(\epsilon)\right\rangle\left\langle u_g'^{\bar{k}}(\epsilon)\right| + Q^{\overline{k-1}}(\epsilon)$, where note that only the decomposition depends on ϵ ($Q^{\bar{k}}$ solves $\underline{\mathrm{X}}^{\bar{k}}(\epsilon)$ but doesn't depend on ϵ). Taking the limit (and using the correct normals) we obtain Equation (7.6).

□

Consider $H > 0$, $G > 0$. Then $H \geq OGO^T$ is equivalent to $H^{-1} \leq OG^{-1}O^T$. For monomial assignments we start higher up in the power diagram (see, e.g., Figure 7.1 and Figure 7.4) and so at some point, we must go down. This corresponds to considering the latter as our matrix instance. Below, we formalise this procedure for later use.

[6] subtleties about degeneracies in $\left|t_h^{\bar{k}}\right\rangle$ are not hard to handle; see Subsection 6.3.3.2, Chapter 6.

[7] It is not hard to see why $H^{\overline{k-1}}$ does not diverge as ϵ goes to zero (granted there was only one diverging eigenvalue in $H^{\bar{k}}$ to start with). The idea is simply to use the reverse Weingarten map; this suppresses the divergence into zero, then one projects out a rank-one subspace. If there was only one zero eigenvalue and if the subspace includes this eigenspace (spanned by a single eigenvector), then the resulting matrix would not have any zero eigenvalues. This can then be inverted to obtain the Weingarten map which is now finite and well-defined.

Definition 161 (Flip Map). Consider an extended matrix instance $\underline{\mathrm{M}}^{\bar{n}} =: \left(H, G, |w\rangle, |v\rangle, H^{\dashv}, G^{\dashv}, |u_h\rangle, |u_g\rangle\right)$. We define the *Flip Map* $\mathscr{F} : \mathbb{M}^n \to \mathbb{M}^n$ as $\underline{\mathrm{M}}^{\bar{n}} \mapsto \left(H^{\dashv}, G^{\dashv}, |w\rangle, |v\rangle, H, G, |u_h\rangle, |u_g\rangle\right) =: \mathscr{F}(M^{\bar{n}})$.

We have introduced all the notation and the main tools that we need. We can now state the solution to the unbalanced f_0-assignment formally, as promised.

§ 7.5 f_0 Unitary (cont.)

7.5.1 The Unbalanced Case

Proposition 162 (The unbalanced f_0 Solution). *Let $t = h - g = \sum_{i=1}^{2n-1} p_i [\![x_i]\!]$ be Mochon's f_0 assignment for the set of real numbers $0 \leq x_1 < x_2 \cdots < x_{2n-1}$. Let $h = \sum_{i=1}^{n-1} p_{h_i} [\![x_{h_i}]\!]$, $g = \sum_{i=1}^{n} p_{g_i} [\![x_{g_i}]\!]$ where p_{h_i} and p_{g_i} are strictly positive, and $\{x_{h_i}\}$ and $\{x_{g_i}\}$ are all distinct. Consider the matrix instance $\underline{X}$:=$(X_h, X_g, |w\rangle, |v\rangle)$ where $X_h \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_{n-1}}, 1/\epsilon)$, $X_g \doteq diag(x_{g_1}, x_{g_2} \ldots x_{g_{n-1}}, x_{g_n})$, $|w\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_{n-1}}}, 0)^T$, $|v\rangle \doteq \left(\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_{n-1}}}, \sqrt{p_{g_n}}\right)^T$. In the limit of $\epsilon \to 0$, the orthogonal matrix*

$$O = \sum_{k=1}^{n} \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right|$$

solves $\underline{X} =: \underline{X}^{\bar{n}}$ (see Definition 155) where the Weingarten Iteration Map (see Definition 157) is used to evaluate $\underline{X}^{\bar{k-1}} = \mathscr{W}(\underline{X}^{\bar{k}})$ until $k = 2$, starting from $k = n$. The Normal Initialisation Map (see Definition 156) is used until $k = 3$ to obtain $\left|u_h^{\bar{k}}\right\rangle$ and $\left|u_g^{\bar{k}}\right\rangle$, viz. $\mathscr{U}(\underline{X}^{\bar{k}}) =: \left(*, \cdots *, \left|u_h^{\bar{k}}\right\rangle, \left|u_g^{\bar{k}}\right\rangle\right)$. The Wiggle-w Normal Initialisation Map (see Definition 156) is used to evaluate $\left|u_h^{\bar{2}}\right\rangle$ and $\left|u_g^{\bar{2}}\right\rangle$, viz. $\mathscr{U}_w(\underline{X}^{\bar{2}}) =: \left(*, *, \left|w^{\bar{2}}\right\rangle, \left|v^{\bar{2}}\right\rangle\right) \oplus \left(*, *, \left|u_h^{\bar{2}}\right\rangle, \left|u_g^{\bar{2}}\right\rangle\right)$. Finally, $\left|u_h^{\bar{1}}\right\rangle := \left|e\left(\left|u_h^{\bar{2}}\right\rangle, \left|w^{\bar{2}}\right\rangle\right)\right\rangle$ and $\left|u_g^{\bar{1}}\right\rangle := \left|e\left(\left|u_g^{\bar{2}}\right\rangle, \left|v^{\bar{2}}\right\rangle\right)\right\rangle$.

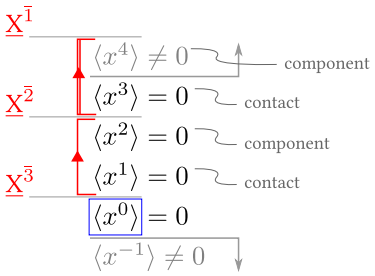


Figure 7.3: Power Diagram representative of an unbalanced f_0 assignment with 5 points (again $n = 3$). Starting upwards from $\langle x^0 \rangle$, one iteration is completed before encountering the instance where the contact condition still holds but the normals do not match, thus the wiggle-w method is employed.

Proof. The argument is essentially the same as that for the balanced case until the very last step. After iterating for l steps, suppose the matrix instance one obtains is $\underline{\mathrm{X}}^{\overline{n-l}}$. To check if another Weingarten iteration is possible, we must check if

$$\left\langle w^{\overline{n-l}}\right| \left(H^{\overline{n-l}}\right)^m \left|w^{\overline{n-l}}\right\rangle - \left\langle v^{\overline{n-l}}\right| \left(G^{\overline{n-l}}\right)^m \left|v^{\overline{n-l}}\right\rangle =$$
$$r\left(\left\langle (H^{\bar{n}})^m\right\rangle, \left\langle (H^{\bar{n}})^{m+1}\right\rangle \ldots, \left\langle (H^{\bar{n}})^{2l+m}\right\rangle\right) - r\left(\left\langle (G^{\bar{n}})^m\right\rangle, \left\langle (G^{\bar{n}})^{m+1}\right\rangle \ldots, \left\langle (G^{\bar{n}})^{2l+m}\right\rangle\right) \tag{7.7}$$

vanishes for both $m = 1$ and $m = 2$, viz.

$$\left\langle x^{2l+1} \right\rangle = 0, \left\langle x^{2l+2} \right\rangle = 0 \tag{7.8}$$

and their lower power analogues (see Equation (7.4)). The $m = 1$ case is the contact condition and $m = 2$ is the component condition (see Definition 155). If $2l+2 \leq 2n-3$ then from Lemma 81 (we use $2n-1$ instead of n in the lemma) it follows that both terms (in Equation (7.7)) become identical and hence the difference indeed vanishes. Consequently, until $l = n-3$ (included), one can iterate to obtain $\underline{\mathrm{X}}^{\bar{n}}, \underline{\mathrm{X}}^{\overline{n-1}}, \ldots \underline{\mathrm{X}}^{\bar{3}}, \underline{\mathrm{X}}^{\bar{2}}$ which in turn can be used to determine $\left|u_h^{\bar{n}}\right\rangle, \left|u_h^{\overline{n-1}}\right\rangle, \ldots, \left|u_h^{\bar{3}}\right\rangle$ and similarly $\left|u_g^{\bar{n}}\right\rangle, \left|u_g^{\overline{n-1}}\right\rangle, \ldots, \left|u_g^{\bar{3}}\right\rangle$ (see Definition 156). Since $\left\langle x^{2n-3=2(n-2)+1} \right\rangle = 0$ but $\left\langle x^{2n-2=2(n-2)+2} \right\rangle \neq 0$ (essentially Equation (7.8) with $l = n-2$), we can use Definition 158 on $\underline{\mathrm{X}}^{\overline{2=n-(n-2)}}$ to determine $\left|u_h^{\bar{2}}\right\rangle$ and $\left|u_g^{\bar{2}}\right\rangle$. The vectors $\left|w^{\bar{1}}\right\rangle$ and $\left|v^{\bar{1}}\right\rangle$ are fixed by the requirement that O is orthogonal and that $O\left|v\right\rangle = \left|w\right\rangle$. As before, if we start with assuming (which we can, see Subsection D.4) that O solves the matrix instance $\underline{\mathrm{X}}^{\bar{n}}$, then using Lemma 152 (and towards the end Lemma 160), we completely determine $O = \sum_{k=1}^{n} \left|u_h^{\bar{k}}\right\rangle \left\langle u_g^{\bar{k}}\right|$. □

The argument can again be concisely represented using a diagram (see Figure 7.3). For concreteness, set $n = 3$ in which case, we must use a wiggle-v step at $\underline{\mathrm{X}}^{\bar{2}}$ which is represented by a double-lined arrow from $\left\langle x^3 \right\rangle$ to $\left\langle x^4 \right\rangle$. As we shall see, this argument can be extended to work with monomial assignments as well. The difference is that we start not at the bottom of the diagram, but higher up, depending on the order of the monomial.

§ 7.6 m Unitary | Solution to Mochon's Monomial Assignments

7.6.1 Simplest Monomial Problem

Recall that the f_0-assignment corresponded to starting at the bottom of the diagram, i.e. at $\left\langle x^0 \right\rangle$ (see Section 7.2). We now consider the simplest monomial problem which corresponds to starting at the top of the diagram, i.e. at $\left\langle x^{2n-2} \right\rangle$ (explained below). Intuitively, while earlier every iteration was leading to an increase in the power of x (in terms of the form $\left\langle x^k \right\rangle$), here every iteration leads to a decrease in the power. This is because we start with inverting the matrices. Later, we use a combination of these strategies to construct the solution.

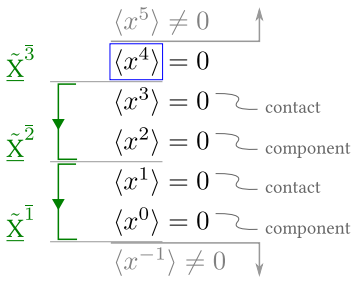


Figure 7.4: Power diagram representative of the simplest monomial assignment for $2n = 6$ points.

Example 163 (Solving the Simplest Monomial Problem.)**.** Suppose the assignment we wish to solve is

$$t = \sum_{i=1}^{2n} -\frac{(-x_i)^{2n-2}}{\prod_{j \neq i}(x_j - x_i)} [\![x_i]\!] = \sum_{i=1}^{2n} \tilde{p}_i [\![x_i]\!]$$

where $0 < x_1 < x_2 \cdots < x_n$. This can be solved using the f_0-solution (see Proposition 153) by writing $t = \sum_{i=1}^{2n} \frac{1}{\prod_{j\neq i}(\omega_j - \omega_i)} [\![x_i]\!]$, where $\omega_i = 1/x_i$, which is in turn equivalent to solving $t' = \sum_{i=1}^{2n} -\frac{1}{\prod_{j\neq i}(\omega_j-\omega_i)} [\![\omega_i]\!]$ (see Corollary 190 with $k = 0$). Instead, we solve this problem using another method—we use $X^{\dashv}$s instead of Xs as in the usual f_0 solution and the fact that $\sum_i \tilde{p}_i x_i^{-k} = 0$ for $k \leq 2n-2$ (see Lemma 81). Let us write t as

$$t = \sum_{i=1}^{n} \tilde{p}_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} \tilde{p}_{g_i} [\![x_{g_i}]\!] = \sum_{i=1}^{n} x_{h_i}^{2n-2} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^{2n-2} p_{g_i} [\![x_{g_i}]\!] .$$

Let the matrix instance corresponding to t be given by $\underline{X}^{\bar{n}} := \left(X_h^{\bar{n}}, X_g^{\bar{n}}, (X_h^{\bar{n}})^{n-1} \left|w^{\bar{n}}\right\rangle, (X_g^{\bar{n}})^{n-1} \left|v^{\bar{n}}\right\rangle\right)$, where

$$X_h^{\bar{n}} \doteq \mathrm{diag}(x_{h_1}, x_{h_2} \ldots x_{h_n}), \qquad X_g^{\bar{n}} \doteq \mathrm{diag}(x_{g_1}, x_{g_2} \ldots x_{g_n}),$$
$$\left|w^{\bar{n}}\right\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_n}}), \qquad \left|v^{\bar{n}}\right\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_n}}).$$

Solving the matrix instance $\underline{X}^{\bar{n}}$ requires us to find an orthogonal matrix O such that $X_h^{\bar{n}} \geq O X_g^{\bar{n}} O^T$ and $O(X_g^{\bar{n}})^{n-1} \left|v^{\bar{n}}\right\rangle = (X_h^{\bar{n}})^{n-1} \left|w^{\bar{n}}\right\rangle$. The matrix inequality can be equivalently written as $\tilde{X}_h^{\bar{n}} \leq O \tilde{X}_g^{\bar{n}} O^T$ where $\tilde{X}_h^{\bar{n}} = (X_h^{\bar{n}})^{-1}$ and $\tilde{X}_g^{\bar{n}} = (X_g^{\bar{n}})^{-1}$. Note that under a change of the direction of the matrix inequality the arguments used in the proof of Lemma 152 go through unchanged. We can therefore consider the matrix instance $\underline{\tilde{X}}^{\bar{n}} := \left(\tilde{X}_h^{\bar{n}}, \tilde{X}_g^{\bar{n}}, \left|\tilde{w}^{\bar{n}}\right\rangle, \left|\tilde{v}^{\bar{n}}\right\rangle\right)$ where $\left|\tilde{w}^{\bar{n}}\right\rangle := (X_h^{\bar{n}})^{n-1} \left|w^{\bar{n}}\right\rangle$ and $\left|\tilde{v}^{\bar{n}}\right\rangle := (X_g^{\bar{n}})^{n-1} \left|v^{\bar{n}}\right\rangle$. After iterating for l steps, suppose the matrix instance one obtains is $\underline{\tilde{X}}^{\overline{n-l}}$. To check if another isometric iteration is possible, we must check if the contact condition (see Definition 155) holds, i.e. if

$$\begin{aligned}
&\left\langle \tilde{w}^{\overline{n-l}} \right| \tilde{H}^{\overline{n-l}} \left| \tilde{w}^{\overline{n-l}} \right\rangle - \left\langle \tilde{v}^{\overline{n-l}} \right| \tilde{G}^{\overline{n-l}} \left| \tilde{v}^{\overline{n-l}} \right\rangle \\
&= r\left(\left\langle \tilde{w}^{\bar{n}} \right| (\tilde{X}_h^{\bar{n}})^1 \left| \tilde{w}^{\bar{n}} \right\rangle, \left\langle \tilde{w}^{\bar{n}} \right| (\tilde{X}_h^{\bar{n}})^2 \left| \tilde{w}^{\bar{n}} \right\rangle \ldots, \left\langle \tilde{w}^{\bar{n}} \right| (\tilde{X}_h^{\bar{n}})^{2l+1} \left| \tilde{w}^{\bar{n}} \right\rangle \right) \\
&\quad - r\left(\left\langle \tilde{v}^{\bar{n}} \right| (\tilde{X}_g^{\bar{n}})^1 \left| \tilde{v}^{\bar{n}} \right\rangle, \left\langle \tilde{v}^{\bar{n}} \right| (\tilde{X}_g^{\bar{n}})^2 \left| \tilde{v}^{\bar{n}} \right\rangle \ldots, \left\langle \tilde{v}^{\bar{n}} \right| (\tilde{X}_g^{\bar{n}})^{2l+1} \left| \tilde{v}^{\bar{n}} \right\rangle \right) \\
&= r\left(\left\langle (X_h^{\bar{n}})^{2n-3} \right\rangle, \left\langle (X_h^{\bar{n}})^{2n-4} \right\rangle \ldots, \left\langle (X_h^{\bar{n}})^{2n-2l-3} \right\rangle \right) \\
&\quad - r\left(\left\langle (X_g^{\bar{n}})^{2n-3} \right\rangle, \left\langle (X_g^{\bar{n}})^{2n-4} \right\rangle \ldots, \left\langle (X_g^{\bar{n}})^{2n-2l-3} \right\rangle \right)
\end{aligned}$$

vanishes. We used Lemma 154 (with $m = 1$) to obtain the RHS (and we continue using the convention that $\left\langle (X_h^{\bar{n}})^k \right\rangle = \left\langle w^{\bar{n}} \right| (X_h^{\bar{n}})^k \left| w^{\bar{n}} \right\rangle$ and similarly $\left\langle (X_g^{\bar{n}})^k \right\rangle = \left\langle v^{\bar{n}} \right| (X_g^{\bar{n}})^k \left| v^{\bar{n}} \right\rangle$). Recall that (see Equation (7.4))

$$\left\langle (H^{\bar{n}})^k \right\rangle - \left\langle (G^{\bar{n}})^k \right\rangle = \left\langle x^k \right\rangle . \tag{7.9}$$

If $0 \leq 2n-2l-3 \leq 2n-2$ then from Lemma 81 it follows that both terms become identical and hence the difference indeed vanishes (one can similarly verify the component condition). Hence, until $l = n-2$ (included), one can apply the Weingarten Iteration to obtain $\left|\tilde{u}_h^{\bar{n}}\right\rangle, \left|\tilde{u}_h^{\overline{n-1}}\right\rangle, \ldots, \left|\tilde{u}_h^{\overline{n-l}}\right\rangle, \ldots, \left|\tilde{u}_h^{\bar{1}}\right\rangle$ and $\left|\tilde{u}_g^{\bar{n}}\right\rangle, \left|\tilde{u}_g^{\overline{n-1}}\right\rangle, \ldots, \left|\tilde{u}_g^{\overline{n-l}}\right\rangle, \ldots, \left|\tilde{u}_g^{\bar{1}}\right\rangle$, which completely determine $O = \sum_{i=1}^{n} \left|\tilde{u}_h^{i}\right\rangle \left\langle \tilde{u}_g^{i}\right|$. The argument can, as before, be concisely represented using a diagram (see Figure 7.4).

7.6.2 Balanced Monomial Problem

Before we start mixing the two approaches, we state a result which helps us keep track of the powers which appear in the contact and component conditions of matrix instances, after we have made a certain number of iterations in both directions.

Lemma 164 (Up-then-Down Contact/Component Lemma)**.** *Consider the extended matrix instance*

$$\underline{M}'^{\bar{n}} := \mathscr{U}(H'^{\bar{n}}, G'^{\bar{n}}, \left|w'^{\bar{n}}\right\rangle, \left|v'^{\bar{n}}\right\rangle, (H'^{\bar{n}})^{\dashv}, (G'^{\bar{n}})^{\dashv}, |.\rangle, |.\rangle).$$

Suppose the Normal Initialisation Map and the Weingarten Iteration Map (see Definition 156 and Definition 157) are applied k times to obtain $\underline{M}'^{\overline{n-k}}$. Let $n-k=d$ and consider $\underline{\tilde{M}}^{\bar{d}} = \mathscr{U}(\mathscr{F}(\underline{M}'^{\bar{d}}))$. Suppose the Normal Initialisation Map and the Weingarten Iteration map are applied l more times to obtain $\underline{\tilde{M}}^{\overline{d-l}} =: \left(\tilde{H}^{\overline{d-l}}, \tilde{G}^{\overline{d-l}}, \left|\tilde{w}^{\overline{d-l}}\right\rangle, \left|\tilde{v}^{\overline{d-l}}\right\rangle, *, \cdots *\right)$. Then,

$$\left\langle \tilde{v}^{\overline{n-k-l}} \right| \left(\tilde{G}^{\overline{n-k-l}}\right)^{\mu} \left| \tilde{v}^{\overline{n-k-l}} \right\rangle = r\left(\left\langle (G'^{\bar{n}})^{-(2l+\mu)} \right\rangle, \ldots, \left\langle (G'^{\bar{n}})^{2k-1+\mu} \right\rangle, \left\langle (G'^{\bar{n}})^{2k+\mu} \right\rangle\right)$$

where $\mu \geq 1$ and r is a multi-variate function which does not have an implicit dependence on $\left\langle (G'^{\bar{n}})^{i} \right\rangle := \left\langle v'^{\bar{n}} \right| (G'^{\bar{n}})^{i} \left| v'^{\bar{n}} \right\rangle$ for any i. The corresponding statement involving H's and $|w\rangle$'s also holds.

This can be proved by combining Lemma 194, Lemma 195 and Lemma 196 (see Section D.5 of the Appendix).

A monomial problem can either be balanced or unbalanced (see Definition 80). We find the solution in these two cases separately, starting with the former. Recall that if a solution requires $\epsilon \to 0$, it does not correspond to anything unphysical (see the argument before Proposition 162).

Proposition 165 (Solving the Balanced Monomial Problem)**.** *Let*

$$t = \sum_{i=1}^{2n} -\frac{(-x_i)^m}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!] = \sum_{i=1}^{n} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!]$$

be a balanced monomial assignment for the set of real numbers $0 < x_1 < x_2 \cdots < x_{2n-1} < x_{2n}$ (see Definition 80; it enforces $0 \leq m \leq 2n-2$) where p_{h_i} and p_{g_i} are strictly positive and $\{x_{h_i}\}$ and $\{x_{g_i}\}$ are all distinct. Note that for both $m=0$ and $m=2n-2$ the problem reduces to the f_0-assignment (see Proposition 162) using Corollary 190 in the latter case. For the remaining cases, consider the corresponding matrix instance $\underline{X}^{\bar{\eta}} := (X_h^{\bar{\eta}}, X_g^{\bar{\eta}}, (X_h^{\bar{\eta}})^b |w\rangle, (X_g^{\bar{\eta}})^b |v\rangle)$ where

- *if $b = m/2$ is an integer (the* aligned case*) then $\eta = n$, $j' = j = 1$,*

$$X_h^{\bar{n}} \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_n}), \qquad X_g^{\bar{n}} \doteq diag(x_{g_1}, x_{g_2} \ldots x_{g_n}),$$
$$\left|w^{\bar{n}}\right\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_n}}), \qquad \left|v^{\bar{n}}\right\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_n}}).$$

- *else if $b = m/2$ is not an integer (the* misaligned case*) then $\eta = n+1$, $j' = 3$, $j = 4$,*

$$X_h^{\overline{n+1}} \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_n}, 1/\epsilon), \qquad X_g^{\overline{n+1}} \doteq diag(x_{g_1}, x_{g_2} \ldots x_{g_n}, \epsilon),$$
$$\left|w^{\overline{n+1}}\right\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_n}}, 0), \qquad \left|v^{\overline{n+1}}\right\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_n}}, 0).$$

Let $k = \left\lfloor \frac{2n-2-m}{2} \right\rfloor$. In the limit of $\epsilon \to 0$, the matrix instance is solved by (note that the sums run backwards)

$$O = \sum_{i=\eta}^{\eta-k+1} \left|u_h'^{i}\right\rangle \left\langle u_g'^{i}\right| + \sum_{i=\eta-k}^{j} \left|\tilde{u}_h^{i}\right\rangle \left\langle \tilde{u}_g^{i}\right| + (1-\delta_{j,j'}) \sum_{i=j'}^{1} \left|u_h'^{i}\right\rangle \left\langle u_g'^{i}\right|,$$

where the terms of the first sum are evaluated in the same way for both cases (i.e. regardless of the alignment). We start with $\underline{M}'^{\bar{\eta}} := \mathscr{U}\left(\underline{X}^{\bar{\eta}} \oplus \left((X_h^{\bar{\eta}})^{-1}, (X_g^{\bar{\eta}})^{-1}, |.\rangle, |.\rangle\right)\right)$ (see Definition 155, Definition 156, Definition 157) and we define

$$\underline{M}'^{l} =: \left(*, \cdots *, \left|u_h'^{l}\right\rangle, \left|u_h'^{l}\right\rangle\right) \qquad \eta - k + 1 \leq l \leq \eta$$

using the relations

$$\underline{M}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{M}^{\dot{l}})) \qquad\qquad \eta - k + 1 \leq l - 1 \leq \eta - 1.$$

The terms of the second sum are also the same in both cases. We start with

$$\underline{\tilde{M}}^{\overline{\eta-k}} := \mathscr{U}(\mathscr{F}(\underline{M}^{\overline{\eta-k}}))$$

and using the relations

$$\underline{\tilde{M}}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{\tilde{M}}^{\dot{l}})) \qquad\qquad j' \leq l - 1 \leq \eta - k - 1$$

we define

$$\left(*, \cdots *, \left|\tilde{u}_h^{\bar{l}}\right\rangle, \left|\tilde{u}_g^{\dot{l}}\right\rangle\right) := \underline{\tilde{M}}^{\dot{l}} \qquad\qquad j \leq l \leq \eta - k.$$

At this point, the aligned problem is solved.

We use the following relations to specify the terms of the third sum, (which solves the misaligned problem):

$$\underline{\tilde{M}}^{\bar{3}} := \mathscr{U}_v(\mathscr{W}(\underline{\tilde{M}}^{\bar{4}}))$$
$$\underline{M}'^{\bar{2}} := \mathscr{U}_w(\mathscr{F}(\mathscr{W}_v(\underline{\tilde{M}}^{\bar{3}}))) =: \left(*, *, \left|w'^{\bar{2}}\right\rangle, \left|v'^{\bar{2}}\right\rangle, *, *, \left|u_h'^{\bar{2}}\right\rangle, \left|u_g'^{\bar{2}}\right\rangle\right)$$
$$\left|u_h'^{\bar{1}}\right\rangle := \left|e\left(\left|u_h'^{\bar{2}}\right\rangle, \left|w'^{\bar{2}}\right\rangle\right)\right\rangle$$

and

$$\left|u_g'^{\bar{1}}\right\rangle := \left|e\left(\left|u_g'^{\bar{2}}\right\rangle, \left|v'^{\bar{2}}\right\rangle\right)\right\rangle,$$

where we used Definition 161, Definition 158, Definition 159.

Proof. We first prove that O solves $\underline{X}^{\bar{n}}$ in the *aligned case* (i.e. when $b = m/2$ is an integer; see Figure 7.5 and note that $\eta = n$ in this case). We denote the components of $\underline{M}'^{\dot{l}}$ by

$$\left(H'^{\dot{l}}, G'^{\dot{l}}, \left|w'^{\dot{l}}\right\rangle, \left|v'^{\dot{l}}\right\rangle, * \ldots, *\right) := \underline{M}'^{\dot{l}}.$$

To start with, we check if $\underline{M}'^{\bar{n}}$ satisfies the contact condition, which corresponds to

$$\left\langle w'^{\bar{n}}\right| H'^{\bar{n}} \left|w'^{\bar{n}}\right\rangle = \left\langle v'^{\bar{n}}\right| G'^{\bar{n}} \left|v'^{\bar{n}}\right\rangle.$$

The LHS is simply $\left\langle w^{\bar{n}}\right| (X_h^{\bar{n}})^{2b+1} \left|w^{\bar{n}}\right\rangle = \left\langle (X_h^{\bar{n}})^{m+1}\right\rangle$ and similarly the RHS is $\left\langle (X_g^{\bar{n}})^{m+1}\right\rangle$. The condition can then be expressed as $\left\langle x^{m+1}\right\rangle = 0$. The component condition similarly can be expressed as $\left\langle x^{m+2}\right\rangle = 0$. From Lemma 81, we know that these conditions hold for $m + 2 \leq 2n - 2$, i.e. $m \leq 2n - 4$ (see Figure 7.5 with $2n = 10$, which means that m can be at most 6 for the contact/component conditions to hold). Assuming $m \leq 2n-4$ we[8] can apply the Weingarten Iteration Map (Definition 157) and use Lemma 152 along with the Normal Initialisation Map (see Definition 156) to construct a part of the solution, viz. use $\underline{M}'^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{M}'^{\dot{l}}))$. Suppose we iterate κ times to obtain $\underline{M}'^{\overline{n-\kappa}}$ (note that κ and k are distinct symbols). The contact condition now corresponds to

$$\left\langle w'^{\overline{n-\kappa}}\right| H'^{\overline{n-\kappa}} \left|w'^{\overline{n-\kappa}}\right\rangle = \left\langle v'^{\overline{n-\kappa}}\right| G'^{\overline{n-\kappa}} \left|v'^{\overline{n-\kappa}}\right\rangle.$$

[8] The $m = 2n - 3$ case can't arise here by the alignment assumption; the $m = 2n - 2$ case becomes a special case which we have seen already—the simplest monomial assignment (see Example 163).

The RHS can be written as

$$r\left(\left\langle w'^{\bar{n}}\right|\left(H'^{\bar{n}}\right)^{1}\left|w'^{\bar{n}}\right\rangle,\left\langle w'^{\bar{n}}\right|\left(H'^{\bar{n}}\right)^{2}\left|w'^{\bar{n}}\right\rangle\ldots\left\langle w'^{\bar{n}}\right|\left(H'^{\bar{n}}\right)^{2\kappa+1}\left|w'^{\bar{n}}\right\rangle\right)$$

using Lemma 154. Similarly for the LHS. The contact condition can then be expressed as $\left\langle x^{2\kappa+1+m}\right\rangle = 0$ (the lower power terms also satisfy this condition if the highest power term does). Proceeding similarly, the component condition can be expressed as $\left\langle x^{2\kappa+2+m}\right\rangle = 0$. From Lemma 81, we know that these conditions hold if $2\kappa+2+m \leq 2n-2$ which yields $\kappa \leq n-b-2 = k-1$. Hence, we can deduce that if O solves the matrix instance then it must have the form $O = \sum_{l=1}^{n-k+1}\left|u_h'^{l}\right\rangle\left\langle u_g'^{l}\right| + Q^{\overline{n-k}}$ where $Q^{\overline{n-k}}$ is an isometry acting on the orthogonal space which remains to be determined. To proceed, we can apply the Weingarten Iteration Map to $\underline{\mathrm{M}}'^{\overline{n-k+1}}$ and obtain $\mathscr{W}(\underline{\mathrm{M}}'^{\overline{n-k+1}}) =: \underline{\mathrm{M}}'^{\overline{n-k}}$, but this instance satisfies neither the contact nor the component condition (corresponds to $\underline{\mathrm{M}}'^{\bar{3}}$ in Figure 7.5). This can be remedied by proceeding as in Example 163.

For this paragraph, let $\left(H, G, |w\rangle, |v\rangle, H^{\dashv}, G^{\dashv}, *, *\right) := \underline{\mathrm{M}}'^{\overline{n-k}}$. Solving $\underline{\mathrm{M}}'^{\overline{n-k}}$ corresponds to finding a Q such that $Q|v\rangle = |w\rangle$ and $H \geq QGQ^{T}$. The matrix inequality can equivalently be written as $H^{\dashv} \leq QG^{\dashv}Q^{T}$. Intuitively, using H and G to evaluate the normals led to contact/component conditions which correspond to increasing powers in the condition $\left\langle x^{l}\right\rangle = 0$. Using $H^{\dashv}$ and $G^{\dashv}$ should decrease the powers and thereby allow us to proceed. We formalise this and use Lemma 164 to bolster the intuition.

We evaluate

$$\tilde{\underline{\mathrm{M}}}^{\overline{n-k}} = \mathscr{U}(\mathscr{F}(\mathscr{W}(\underline{\mathrm{M}}'^{\overline{n-k+1}}))$$

and let $\tilde{\underline{\mathrm{M}}}^{\dot{l}} =: \left(\tilde{H}^{\dot{l}}, \tilde{G}^{\dot{l}}, \left|\tilde{w}^{\dot{l}}\right\rangle, \left|\tilde{v}^{\dot{l}}\right\rangle\right)$ (this step is indicated by the small triangles next to $\underline{\mathrm{M}}'^{\bar{3}}$ and $\tilde{\underline{\mathrm{M}}}^{\bar{3}}$ in Figure 7.5). Let the matrix instance one obtains after iterating l times using $\tilde{\underline{\mathrm{M}}}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\tilde{\underline{\mathrm{M}}}^{\dot{l}}))$ starting with $\tilde{\underline{\mathrm{M}}}^{\overline{n-k}}$ be $\tilde{\underline{\mathrm{M}}}^{\overline{n-k-l}}$. The contact condition for $\tilde{\underline{\mathrm{M}}}^{\overline{n-k-l}}$ is

$$\left\langle \tilde{w}^{\overline{n-k-l}}\right| \tilde{H}^{\overline{n-k-l}} \left|\tilde{w}^{\overline{n-k-l}}\right\rangle = \left\langle \tilde{v}^{\overline{n-k-l}}\right| \tilde{G}^{\overline{n-k-l}} \left|\tilde{v}^{\overline{n-k-l}}\right\rangle,$$

which effectively becomes $\left\langle x^{-(2l+1)+m}\right\rangle = 0$ using Lemma 164, noting that the lowest power is relevant here, and that $\left|w'^{\bar{n}}\right\rangle = (X_h^{\bar{n}})^{m/2}\left|w^{\bar{n}}\right\rangle$ (similarly for $\left|v'^{\bar{n}}\right\rangle$). We can analogously see that the component condition yields $\left\langle x^{-(2l+2)+m}\right\rangle = 0$. From Lemma 81, we know that these conditions hold if $0 \leq -(2l+2)+m$ which yields $l \leq b-1$. This means that the rank, i.e. $n-k-l$, until which the contact/component condition holds is $\cancel{n-n}+1+\cancel{b-b}+1 = 2$ (included) where we used $k = n-b-1$. Hence we deduce that if $Q^{\overline{n-k}}$ resolves $\tilde{\underline{\mathrm{M}}}^{\overline{n-k}}$ then it must have the form $Q^{\bar{k}} = \sum_{l=n-k}^{1}\left|\tilde{u}_h^{l}\right\rangle\left\langle \tilde{u}_g^{l}\right|$ (using Lemma 152) which completely specifies $Q^{\bar{k}}$, proving (together with the previous argument) that O solves $\underline{\mathrm{M}}^{\bar{n}}$.

We now prove that O solves $\underline{\mathrm{X}}^{\overline{n+1}}$ in the misaligned case (i.e. when $m/2$ is not an integer; see Figure 7.5). We can proceed as in the aligned case until the contact/component condition is violated. In this case, after κ steps the said condition is $\left\langle x^{2\kappa+2+m}\right\rangle = 0$ which holds until $2\kappa+2+m \leq 2n-2$ (using Lemma 81). This corresponds to $\kappa \leq \frac{2n-2-m}{2} - 1$, which yields $\kappa \leq k-1$. Hence $\underline{\mathrm{M}}'^{\overline{\eta-k+1}}$ will be the last instance satisfying the required contact/component conditions (this corresponds to $\underline{\mathrm{M}}'^{\bar{5}}$ in Figure 7.5; use $(n+1)-(k-1)$ with $n=5, k=2$). Supposing O solves $\underline{\mathrm{X}}^{\overline{n+1}}$ we deduce (using Lemma 152 and the arguments from the previous case) that it must have the form $O = \sum_{l=\eta}^{\eta-k+1}\left|u_h'^{l}\right\rangle\left\langle u_g'^{l}\right| + Q^{\overline{\eta-k}}$. At the instance $\underline{\mathrm{M}}'^{\overline{\eta-k}} = \mathscr{W}(\underline{\mathrm{M}}'^{\eta-k+1})$ we flip as before to obtain $\tilde{\underline{\mathrm{M}}}^{\overline{\eta-k}} = \mathscr{U}(\mathscr{F}(\underline{\mathrm{M}}'^{\overline{n-k}}))$ (these are indicated by the triangles next to $\underline{\mathrm{M}}'^{\bar{4}}$ and $\tilde{\underline{\mathrm{M}}}^{\bar{4}}$ in Figure 7.5). We proceed as before to write the contact/component condition after l iterations, $\left\langle x^{-(2l+2)+m}\right\rangle = 0$ which from Lemma 81 holds if

$0 \leq -(2l+2)+m$. This in turn yields $l \leq m/2-1$ entailing that the rank, i.e. $\eta-k-l$, until which the contact/component condition holds is $\cancel{\kappa}+1-(\cancel{\kappa}-1+\lfloor -m/2 \rfloor)-(\lfloor m/2 \rfloor -1)=4$ (this corresponds to $\underline{\tilde{\mathrm{M}}}^{\bar{4}}$ in Figure 7.5). Continuing with the argument for the form of O, we can deduce (again, using Lemma 152 and the previous reasoning) that $Q^{\overline{\eta-k}} = \sum_{l=\eta-k}^{4} \left|\tilde{u}_h^l\right\rangle \left\langle \tilde{u}_g^l\right| + Q^{\bar{3}}$. Since $\underline{\tilde{\mathrm{M}}}^{\bar{4}}$ satisfies the required contact/component conditions, we can iterate once more. However, at this point, only the contact condition holds but the component condition does not (see Figure 7.5). Consider $\underline{\tilde{\mathrm{M}}}^{\bar{3}} = \mathscr{U}_v(\mathscr{W}(\underline{\tilde{\mathrm{M}}}^{\bar{4}}))$ and let $(\tilde{H}^{\bar{3}}, \tilde{G}^{\bar{3}}, *, \cdots *) := \underline{\tilde{\mathrm{M}}}^{\bar{3}}$. We can not apply Lemma 152 on $\underline{\tilde{\mathrm{M}}}^{\bar{3}}$ but we can apply Lemma 160 as $\underline{\tilde{\mathrm{M}}}^{\bar{3}}$ has wiggle-v room ϵ along $|n+1\rangle$ (see Definition 155). To see this, note that the probability vectors had no component along $|n+1\rangle$ and that we inverted the matrices using the flip map. This yields $Q^{\bar{3}} = \left|\tilde{u}_h^{\bar{3}}\right\rangle \left\langle \tilde{u}_g^{\bar{3}}\right| + Q^{\bar{2}}$. The lemma also lets us proceed by the application of the Wiggle-v Iteration map (see Definition 159) $\underline{\tilde{\mathrm{M}}}^{\bar{2}} = \mathscr{W}_v(\underline{\tilde{\mathrm{M}}}^{\bar{3}})$. Since at this point even the contact condition does not hold, we again apply the flip map (and the wiggle-w initialisation map as justified next) to obtain $\underline{\mathrm{M}'}^{\bar{2}} = \mathscr{U}_w(\mathscr{F}(\underline{\tilde{\mathrm{M}}}^{\bar{2}}))$. Instead of decreasing the power of x, the contact condition of this instance corresponds to increasing the power of x, i.e. the contact condition for $\underline{\mathrm{M}'}^{\bar{2}}$ corresponds to $\left\langle x^{2(k-1)+2+m+1}\right\rangle = 0$ which in turn holds if $2k+m+1 \leq 2n-2$. Indeed, $0 = \cancel{2n-2} + 2\lfloor -m/2 \rfloor + m + 1 \leq \cancel{2n-2} = 0$ (substituting for $n=5, k=2, m=3$ we get $8 = 2\cdot 2+3+1 \leq 2\cdot 5-2 = 8$). Since $\underline{\mathrm{M}'}^{\bar{2}}$ has wiggle-w room ϵ along $|n+1\rangle$, we were justified at applying the wiggle-w initialisation map (see Lemma 160). This, and the orthogonality of O, determine the form of $Q^{\bar{2}} = \left|u_h'^{\bar{2}}\right\rangle \left\langle u_g'^{\bar{2}}\right| + \left|u_h'^{\bar{1}}\right\rangle \left\langle u_g'^{\bar{1}}\right|$, which in turn completely determines the solution, O.

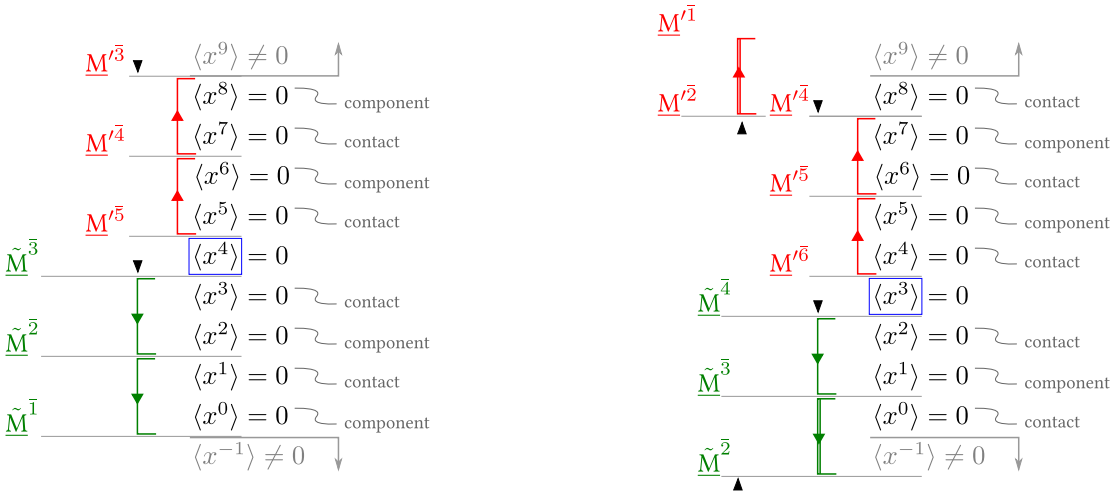


Figure 7.5: Power diagram representative of the aligned (left) and misaligned (right) balanced monomial assignment for $2n=10$ with $m=4$ (left) and $m=3$ (right).

□

7.6.3 Unbalanced Monomial Problem

In the case of an unbalanced monomial problem, either there is a misalignment at the top or at the bottom. If the misalignment is at the top, it is cleaner to start with going downwards. To facilitate the tracking of powers, we state a result similar to Lemma 164, where we start with going downwards.

Lemma 166 (Down-then-Up Contact/Component Lemma)**.** *Consider the matrix instance*

$$\underline{\tilde{M}}^{\bar{n}} := \mathscr{U}\left((H'^{\bar{n}})^{\dashv}, (G'^{\bar{n}})^{\dashv}, \left|w'^{\bar{n}}\right\rangle, \left|v'^{\bar{n}}\right\rangle, H'^{\bar{n}}, G'^{\bar{n}}, |.\rangle, |.\rangle\right).$$

Suppose the Normal Initialisation Map and the Weingarten Iteration Map (see Definition 156 and Definition 157) are applied k times to obtain $\underline{\tilde{M}}^{\overline{n-k}}$. Let $n-k=d$ and consider $\underline{M'}^{\bar{d}} = \mathscr{U}(\mathscr{F}(\underline{\tilde{M}}^{\bar{d}}))$. Suppose the Normal Initialisation Map and the Weingarten Iteration map are applied l more times to obtain $\underline{M'}^{\overline{d-l}} =: \left(H'^{\overline{d-l}}, G'^{\overline{d-l}}, \left|w'^{\overline{d-l}}\right\rangle, \left|v'^{\overline{d-l}}\right\rangle, *, \cdots *\right)$. Then,

$$\left\langle v'^{\overline{n-k-l}}\right| \left(G'^{\overline{n-k-l}}\right)^{\mu} \left|v'^{\overline{n-k-l}}\right\rangle = r\left(\left\langle (G'^{\bar{n}})^{-(2k+\mu)}\right\rangle, \ldots, \left\langle (G'^{\bar{n}})^{2l+\mu-1}\right\rangle, \left\langle (G'^{\bar{n}})^{2l+\mu}\right\rangle\right)$$

where $\mu \geq 1$ and r is a multi-variate function which does not have an implicit dependence on $\left\langle (G'^{\bar{n}})^{i}\right\rangle := \left\langle v'^{\bar{n}}\right| (G'^{\bar{n}})^{i} \left|v'^{\bar{n}}\right\rangle$ for any i. The corresponding statement involving H's and $|w\rangle$'s also holds.

This can be proved by combining Lemma 194, Lemma 195 and Lemma 196 (see Section D.5 of the Appendix).

Finally, we state the solution to the unbalanced monomial problem.

Proposition 167 (Solving the Unbalanced Monomial Problem)**.** *Let*

$$t = \sum_{i=1}^{2n-1} -\frac{(-x_i)^m}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!] = \sum_{i=1}^{n_h} x_{h_i}^m p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} x_{g_i}^m p_{g_i} [\![x_{g_i}]\!]$$

be an unbalanced monomial assignment for the set of real numbers $0 < x_1 < x_2 \cdots < x_{2n-1}$ (see Definition 80) where p_{h_i} and p_{g_i} are strictly positive and $\{x_{h_i}\}$ and $\{x_{g_i}\}$ are all distinct. Note that for both $m=0$ and $m=2n-3$ the problem reduces to the f_0-assignment (see Proposition 162) using Corollary 190 in the latter case. For the remaining cases, consider the corresponding matrix instance $\underline{X}^{\bar{n}} := (X_h^{\overline{n'}}, X_g^{\overline{n'}}, (X_h^{\overline{n'}})^b |w\rangle, (X_g^{\overline{n'}})^b |v\rangle)$ where

- *if $n_h = n$ (the Wiggle-v case; corresponds to odd m)*

$$X_h^{\bar{n}} \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_{n-1}}, x_{h_n}), \qquad X_g^{\bar{n}} \doteq diag(x_{g_1}, x_{g_2} \ldots x_{g_{n-1}}, \epsilon),$$
$$\left|w^{\bar{n}}\right\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_{n-1}}}, \sqrt{p_{h_n}}), \qquad \left|v^{\bar{n}}\right\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_{n-1}}}, 0),$$

- *else if $n_g = n$ (the Wiggle-w case; corresponds to even m)*

$$X_h^{\bar{n}} \doteq diag(x_{h_1}, x_{h_2} \ldots x_{h_{n-1}}, 1/\epsilon), \qquad X_g^{\bar{n}} \doteq diag(x_{g_1}, x_{g_2} \ldots x_{g_{n-1}}, x_{g_n}),$$
$$\left|w^{\bar{n}}\right\rangle \doteq (\sqrt{p_{h_1}}, \sqrt{p_{h_2}} \cdots \sqrt{p_{h_{n-1}}}, 0), \qquad \left|v^{\bar{n}}\right\rangle \doteq (\sqrt{p_{g_1}}, \sqrt{p_{g_2}} \cdots \sqrt{p_{g_{n-1}}}, \sqrt{p_{g_n}}).$$

Consider the Wiggle-v case. Let $k = \frac{2n-3-m}{2}$ (this will be an integer as m is odd). In the limit of $\epsilon \to 0$,

$$O = \sum_{i=n}^{n-k+1} \left|u_h'^{i}\right\rangle \left\langle u_g'^{i}\right| + \sum_{i=n-k}^{1} \left|\tilde{u}_h^{i}\right\rangle \left\langle \tilde{u}_g^{i}\right|$$

solves the matrix instance $\underline{X}^{\bar{n}}$ where the terms in the sum are defined as follows. We start with $\underline{M'}^{\bar{n}} := \mathscr{U}(\underline{X}^{\bar{n}} \oplus \left((X_h^{\bar{n}})^{-1}, (X_g^{\bar{n}})^{-1}, |.\rangle, |.\rangle\right)$ (see Definition 156, Definition 157) and using the relation

$$\underline{M'}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{M'}^{l})) \qquad\qquad n-k+1 \leq l-1 \leq n-1,$$

we define

$$\left(*, \cdots *, \left|u_h'^{l}\right\rangle, \left|u_g'^{l}\right\rangle\right) := \underline{M'}^{l} \qquad\qquad n-k+1 \leq l \leq n.$$

These define the terms of the first sum. For the terms of the second sum we start with

$$\underline{\tilde{M}}^{\overline{n-k}} := \mathscr{U}\left(\mathscr{F}(\underline{M'}^{\overline{n-k}})\right)$$

and using the relation

$$\underline{\tilde{M}}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{\tilde{M}}^{\dot{l}})) \qquad\qquad 3 \leq l-1 \leq n-k-1,$$

we define

$$\left(*, \cdots *, \left|\tilde{u}_h^{\bar{l}}\right\rangle, \left|\tilde{u}_g^{\dot{l}}\right\rangle\right) := \underline{\tilde{M}}^{\dot{l}} \qquad\qquad 2 \leq l \leq n-k.$$

Finally, we define (see Definition 158)

$$\underline{\tilde{M}}^{\bar{\bar{2}}} := \mathscr{U}_v(\mathscr{W}(\underline{\tilde{M}}^{\bar{\bar{3}}})) =: \left(*, *, \left|\tilde{w}^{\bar{\bar{2}}}\right\rangle, \left|\tilde{v}^{\bar{\bar{2}}}\right\rangle, * \cdots *\right),$$

$\left|\tilde{u}_h^{\bar{\bar{1}}}\right\rangle := \left|e\left(\left|\tilde{u}_h^{\bar{\bar{2}}}\right\rangle, \left|\tilde{w}^{\bar{\bar{2}}}\right\rangle\right)\right\rangle$ *and* $\left|\tilde{u}_g^{\bar{\bar{1}}}\right\rangle := \left|e\left(\left|\tilde{u}_g^{\bar{\bar{2}}}\right\rangle, \left|\tilde{v}^{\bar{\bar{2}}}\right\rangle\right)\right\rangle$.

Consider the Wiggle-w case. Let $k = \frac{m}{2}$ *(this will be an integer as* m *is even). In the limit of* $\epsilon \to 0$,

$$O = \sum_{i=n}^{n-k+1} \left|\tilde{u}_h^{\dot{i}}\right\rangle \left\langle \tilde{u}_g^{\dot{i}}\right| + \sum_{i=n-k}^{1} \left|u_h'^{\dot{i}}\right\rangle \left\langle u_g'^{\dot{i}}\right|$$

solves the matrix instance $\underline{X}^{\bar{n}}$ *where the terms in the sum are defined as follows. We start with*

$$\tilde{M}^{\bar{n}} := \mathscr{U}\left(\mathscr{F}\left(\underline{X}^{\bar{n}} \oplus \left((X_h^{\bar{n}})^{-1}, (X_g^{\bar{n}})^{-1}, |.\rangle, |.\rangle\right)\right)\right)$$

(see Definition 156, Definition 161, Definition 157) and using the relation

$$\underline{\tilde{M}}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{\tilde{M}}^{\dot{l}}) \qquad\qquad n-k+1 \leq l-1 \leq n-1,$$

we define

$$\left(*, \cdots *, \left|u_h'^{\dot{l}}\right\rangle, \left|u_g'^{\dot{l}}\right\rangle\right) := \underline{\tilde{M}}^{\dot{l}} \qquad\qquad n-k+1 \leq l \leq n.$$

These determine the terms of the first sum. For the terms of the second sum we start with

$$\underline{M'}^{\overline{n-k}} := \mathscr{U}\left(\mathscr{F}(\underline{\tilde{M}}^{\overline{n-k}})\right)$$

and using

$$\underline{M'}^{\overline{l-1}} := \mathscr{U}(\mathscr{W}(\underline{M'}^{\dot{l}})) \qquad\qquad 3 \leq l-1 \leq n-k-1,$$

we define

$$\left(*, \cdots *, \left|u_h'^{\dot{l}}\right\rangle, \left|u_g'^{\dot{l}}\right\rangle\right) := \underline{M'}^{\dot{l}} \qquad\qquad 2 \leq l \leq n-k.$$

Finally, we define (see Definition 158)

$$\underline{M'}^{\bar{\bar{2}}} := \mathscr{U}_w(\mathscr{W}(\underline{M'}^{\bar{\bar{3}}})) =: (*, *, \left|w'^{\bar{\bar{2}}}\right\rangle, \left|v'^{\bar{\bar{2}}}\right\rangle, * \cdots *),$$

$\left|u_h'^{\bar{\bar{1}}}\right\rangle := \left|e\left(\left|u_h'^{\bar{\bar{2}}}\right\rangle, \left|w'^{\bar{\bar{2}}}\right\rangle\right)\right\rangle$ *and* $\left|u_g'^{\bar{\bar{1}}}\right\rangle := \left|e\left(\left|u_g'^{\bar{\bar{2}}}\right\rangle, \left|v'^{\bar{\bar{2}}}\right\rangle\right)\right\rangle$.

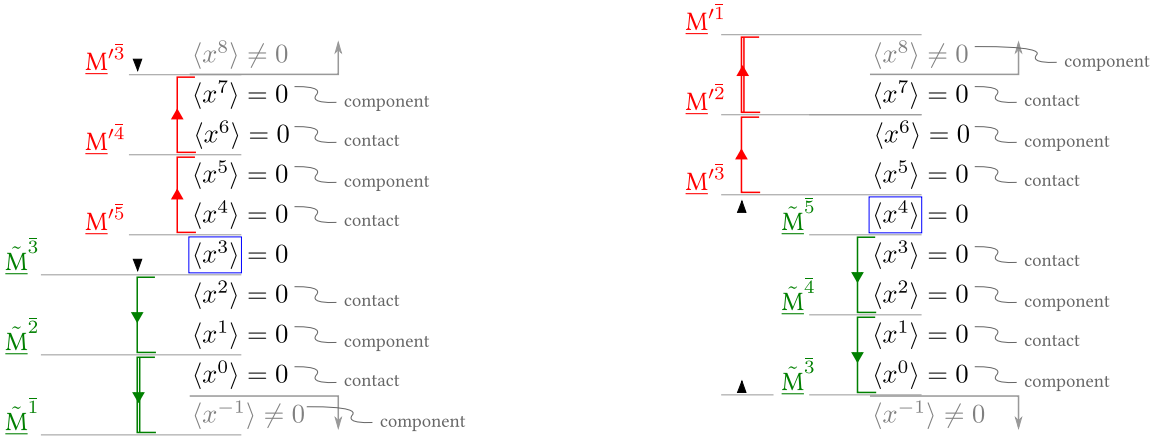


Figure 7.6: Power diagram representative of the unbalanced monomial assignment for $n = 4$ ($2n-1 = 7$) with $m = 3$ (left; wiggle-v case) and $m = 4$ (right; wiggle-w case).

Proof. From Figure 7.6 it is clear that the wiggle-v case is essentially the same as the balanced misaligned monomial until the second to last step (the wiggle-w step after wiggle-v is not needed). From Figure 7.6 it is also clear the wiggle-w case is essentially the same as the wiggle-v case except that we must start with going downwards (decreasing powers of $\langle x^\mu\rangle$), i.e. using $\tilde{\underline{M}}^{\bar{n}}$ and then flip to $\underline{M}'^{\bar{k}}$ to go upwards and end with a wiggle-w iteration. The arguments for the contact/component conditions go through unchanged using Lemma 166. □

Combining the results together, we can now formally state Theorem 11 accompanied by a proof.

Theorem. *Let t be Mochon's f-assignment (see Definition 80) on strictly positive coordinates without loss of generality (see Lemma 83). Suppose f has real and strictly positive roots. Then, decompose the assignment as $t = \sum_i \alpha_i t'_i$ where α_i are positive and t'_i are monomial assignments (see Lemma 82). Each t'_i admits an exact solution[9] of the form given in Proposition 165 or Proposition 167.*

Proof of Theorem 11. From Lemma 83 we can find an f-assignment which has the same solution as the one given, but has all coordinates strictly positive. We therefore consider the latter. From Lemma 82 one can express this assignment as a sum of monomial assignments. A monomial assignment is either balanced—in which case its solution is given by Proposition 165—or it is unbalanced—in which case its solution is given by Proposition 167. □

[9] See *Notation* at the beginning of the Chapter 5.

8

Chapter

Conclusion and Outlook

In Chapter 2 and Chapter 3 we presented the proof of existence of quantum WCF protocol with vanishing bias based on the work of Kitaev/Mochon and Aharonov, Chailloux, Ganz, Kerenidis, and Magnin. In Chapter 4 we described a different way of converting time dependent point games (TDPGs) into explicit protocols, granted a certain matrix subject to a matrix inequality and a vector equality can be found. We used this to construct a protocol approaching bias $1/10$, going below the former best know protocol—the Dip Dip Boom protocol—also due to Mochon, which approached bias $1/6$. In Chapter 5 we gave exact unitaries which correspond to protocols approaching arbitrarily small bias, or more precisely, to $\epsilon_k := 1/(4k+2)$ for any integer $k > 1$. These unitaries suffice to convert Mochon's time independent point games (TIPGs) which he used to prove the existence of protocols with bias ϵ_k into explicit protocols. However, these unitaries are based on Mochon's TIPGs and do not, *a priori*, apply to other TIPGs, such as the one by Pelchat and Høyer [21]. We addressed this in Chapter 6, by introducing what we called the Elliptic Monotone Align (EMA) algorithm which can, in conjunction with the TEF and the prior work, numerically convert any TIPG into an explicit protocol. Numerical algorithms are needed to diagonalise matrices and find the roots of polynomials. In Chapter 7, we again give unitaries corresponding to Mochon's TIPGs by using the ellipsoid picture introduced with the EMA algorithm. The technique has similarities with the EMA algorithm but is different in two key ways—it uses isometries instead of unitaries and relies on exact formulae for the various geometric quantities.

As we have succeeded at describing the construction of quantum WCF protocols with arbitrarily small biases, the more interesting aspects of the problem can now be explored. We describe three possible directions.

(1) Optimality of the construction. Various questions about the optimality of WCF protocols are still unanswered, including that of the conversion process between the frameworks, about Mochon's point games themselves and about the exact explicit WCF protocols.

- *Mochon's Game.* One major shortcoming of the explicit unitaries is that they were obtained by enlarging the Hilbert space,[1] beyond that which is necessary for an optimal solution. One approach towards reducing this could be to understand the connection between the perturbatively defined unitary reported in Chapter 4 and the exact one in Chapter 5, corresponding to the $1/10$-bias protocols. Another approach could be to try and reduce the dimension using a standard technical lemma from Kitaev/Mochon's article (stated as Lemma 104 here). Recall that the Kitaev/Mochon's recipe for converting a TIPG into a TDPG uses what they call a *catalyst state*. For Mochon's game with $\epsilon(1) = 1/6$ bias one can, by inspection, obtain a TDPG which only requires both players to hold qutrits locally and exchange one qubit at each round. However, if one uses the catalyst state based approach, then the dimension of the space scales with the number rounds, which in turn diverges as bias $\epsilon(1) = 1/6$ is approached by the protocol. Can we convert Mochon's TIPG into a TDPG by harnessing the structure of his game? There is a notion of time-ordering at play—any TIPG in which points can be moved about without involving

[1] we expect the dimension of the space to scale exponentially with the number of points involved in the Mochon's f-assignment

causal loops, can be easily converted into a TDPG. The challenge is to formalise this procedure and explore if this can be used to lower bound the bias in order to clarify the reason behind the threshold like behaviour around $\epsilon = 1/6$. By a threshold like behaviour, we mean the following: for $\epsilon = 1/6$ we have a protocol which uses two local qutrits and one message qubit while for $\epsilon < 1/6$ all known protocols need catalysts which correspond to large local registers that scale with the number of points in the associated point-game. These might also have implications on the number of rounds (last point). The recently introduced techniques for proving a lower bound on the number of rounds needed in WCF protocols should be helpful [26].

- *Pelchat-Høyer games.* Another family of TIPGs was proposed by Pelchat and Høyer [21] which achieve arbitrarily low bias. After the verification of this result,[2] it might be interesting to see if a corresponding exact WCF protocol can be obtained; hopefully in fewer dimensions. The Pelchat-Høyer construction is based on considering a combination of valid and invalid basic moves (in the TDPG/TIPG framework) which together form valid non-trivial moves. The challenge is to find a decomposition such that each term stays valid and is still, only slightly non-trivial. The corresponding unitaries could perhaps be determined perturbatively, if they involve a constant number of points unlike the terms in the decomposition of Mochon's assignments. Somewhat complementary to this, it might also be of interest to find the TDPG corresponding to these TIPGs to get some insight into the possible threshold of $\epsilon = 1/6$.
- *Framework.* Constructing general tools to optimise/test the optimality of TIPG for the number of points and the number of rounds in the associated TDPG would be very useful to both constructing better protocols and benchmarking existing ones. The first step towards this end would be, again, to understand if $\epsilon = 1/6$ is indeed a threshold. A less ambitious goal is related to the two general methods ([28] and Chapter 4) which are known for converting a TDPG into an explicit WCF protocol, granted certain unitaries are known. Understanding how they compare and if they are optimal under an appropriately defined notion of optimality is useful. We already know that in terms of the duration of time for which the message register must be kept coherent, the recently introduced method is better. It is also clear that in some cases, even the recent one is sub-optimal as it fails to produce a $1/6$ bias protocol from its TDPG which matches the resource usage of the $1/6$ protocol given by Mochon. An obvious starting point could be to adapt the framework to work better in this particular case.

(2) Relaxing assumptions. The assumptions made to obtain (near) perfect WCF protocols are not realistic.

- *System size.* The size of the incoming system containing the message is assumed to be known, however, this is hard to enforce physically. One way of dropping this assumption could be to adapt the following technique introduced in [20]: imposition of constraints on average energy for prepare and measure like scenarios. The main challenge in our setting would be that the players do not trust each other, however, the tools they developed should prove to be beneficial.
- *Noise.* It is not hard to see that adding noise into a WCF protocol can cause a disagreement even when both players are honest. It has been shown that in the absence of noise but in the present of losses (such as loss of particles during transmission), WCF can still be performed with a certain bias [7]. An interesting open question is: Are there lower bounds to the lossy but noiseless setting? One way of proceeding could be to generalise Kitaev/Mochon's frameworks to handle an additional outcome corresponding to abort and to constraint the unitaries to be such that the cheating player is allowed to control the losses. Even a preliminary understanding of this procedure should allow us to construct protocols with improved bias or at least understand why the existing protocols work, better. One hurdle is the number of rounds which in the loss tolerant protocol can vary depending on the strategy of the malicious player while the Kitaev/Mochon framework is designed for protocols with a constant number of rounds. Here

[2] as, to the best of my knowledge, this was not published in a journal/conference

one should be able to extend the notion of "catalyst state" introduced by Kitaev/Mochon. Next, quantum computation is realistic because error correction allows ideal functionality even with noisy (realistic) hardware. This, however, does not directly mean that WCF can be performed in such a setting. It is not obvious how one could correct errors in the adversarial cryptographic setting without compromising the security. Consider the simplest error correcting code and the simplest WCF protocol. The honest case should work directly but in the case where a malicious player is involved, the evaluation of the bias must involve the communication of the syndrome, decoding the error and finally applying a unitary which corrects for it. These steps can be directly adapted into the Kitaev/Mochon framework with the seemingly minor alteration that a malicious player can influence the unitary of the honest player in a way which is consistent with the noise model. The challenge here would be to make the security claim independent of the noise model. An appropriate relaxation of the constraints in the dual problem might hold the key to this conundrum. Recently, generic techniques have been proposed to study adversarial cryptographic settings [18] which might prove to be the right language for describing such a relaxation. One way of approaching the problem could be to further generalise the technique so that it facilitates the handling of noise without any error correction. This step itself should be of independent value as its results would serve as benchmarks against which error correction based schemes must be compared. A simultaneous but complementary approach could be to construct protocols which are robust against specific models of noise, such as those appearing in quantum optics. The insights from the two approaches should quicken the advance towards the final construction.

- *Device Dependence.* Despite being an adversarial setup, device independent WCF protocols have been suggested which involve the exchange of quantum boxes (analogous to exchanging qubits) [2]. The bias, however, was abysmal and to date, no improvement has been reported and no lower bound on the bias is known. The first step could be to redefine the protocol in a generalisable way; perhaps construct successively worse protocols (by, for instance, using fewer boxes) and subsequently, try to view them as belonging to the same family. One could even consider trying to use PR-boxes or similar non-signalling boxes to understand the behaviour better. A complimentary approach could be to construct the analogue of the Kitaev/Mochon framework where instead of qubits and unitaries, one considers more abstract objects (which simulate the exchange of boxes) which are constrained only by their statistics. Recently, WCF protocols were looked at from a general probability theory (GPT) perspective (which were used to extend the impossibility results theories beyond quantum). They used conic duality more generally which is what the Kitaev/Mochon frameworks were built on and hence, this GPT based approach could be the starting point [36].

(3) A fundamental connection. It is known that perfect WCF implies optimal SCF [12]. Does this work the other way? This question may be more general than quantum because the construction in [12] is purely classical. One way of proceeding could be to try to construct optimal SCF protocols directly by adapting Kitaev/Mochon's technique and using known, simpler protocols as a starting point. The insights might not only help answer the question but might, additionally, yield another construction for perfect WCF.

IV

Part

Appendix

A

Appendix

Existence of almost perfect weak coin flipping

§ A.1 WCF protocol as an SDP and its Dual

Proof (sketch) of Theorem 20. In this proof, we only show that $P_B^* \leq \min \mathrm{tr}(Z_{A,0} \left|\psi_{A,0}\right\rangle \left\langle\psi_{A,0}\right|)$ instead of the equality (and similarly for P_A^*). This is weak duality. The proof of strong duality, i.e. equality, see Appendix B of [28].

We start with the primal formulation for P_B^* as

$$P_B^* = \max_{\{\rho_{AM,i}\}} \mathrm{tr}(\rho_{AM,n}\Pi_A^{(1)})$$

subject to the constraints

$$\begin{aligned}
\mathrm{tr}_M(\rho_{AM,i}) &= \mathrm{tr}_M(E_{A,i}U_{A,i}\rho_{AM,i-1}U_{A,i}^\dagger E_{A,i}) && \text{for } i \text{ odd} \\
\mathrm{tr}_M(\rho_{AM,i}) &= \mathrm{tr}_M(\rho_{AM,i-1}) && \text{for } i \text{ even} \\
\mathrm{tr}_M(\rho_{AM,0}) &= \mathrm{tr}_{MB}\left(\left|\psi_0\right\rangle\left\langle\psi_0\right|\right) = \left|\psi_{A,0}\right\rangle\left\langle\psi_{A,0}\right|
\end{aligned}$$

Consider instead the following function:

$$\begin{aligned}
L(\{\rho_{AM,i}\}_{i=1}^n, \{Z_{A,i}\}_{i=1}^n) = \mathrm{tr}(\rho_{AM,n}\Pi_A^{(1)}) \\
+ \mathrm{tr}_A\left(Z_{A,0}\mathrm{tr}_M\left(\rho_{AM,0} - \left|\psi_{A,0}\right\rangle\left\langle\psi_{A,0}\right| \otimes \mathbb{I}_M\right)\right) \\
+ \sum_{i\in\{1,3,\dots\}} \mathrm{tr}_A(Z_{A,i}\mathrm{tr}_M(\rho_{AM,i} - E_{A,i}U_{A,i}\rho_{AM,i-1}U_{A,i}^\dagger E_{A,i})) \\
+ \sum_{i\in\{2,4,\dots\}} \mathrm{tr}_A(Z_{A,i}\mathrm{tr}_M(\rho_{AM,i} - \rho_{AM,i-1}))
\end{aligned}$$

where $Z_{A,i}$ are hermitian matrices on the $\mathcal{A}$ space. Hermiticity guarantees that scalar obtained after the trace, is real. For brevity, we refer to $L(\{\rho_{AM,i}\}_{i=0}^n, \{Z_{A,i}\}_{i=0}^n)$ as $L(\rho, Z)$. Observe that $P_B^* \leq \max_{\rho\geq 0} L(\rho, Z)$ because of the following two reasons. First, the optimal solution $\{\rho_{AM,i}^\star\}_{i=0}^n$ saturates the inequality and removes the dependence on Z. Second, a maximisation over ρ can only increase the value of $L(\rho, Z)$. As the bound holds for all $\{Z_{A,i}\}_{i=1}^n$, the best bound is obtained by minimising over all $Z_{A,i}$s. We are therefore interested in the function $P_B^* \leq \min_Z \max_{\rho\geq 0} L(\rho, Z)$. To this end, we re-express $L(\rho, Z)$ collecting all the dependence on ρ_i. Why this helps should become clear shortly. We have (supposing n is odd for concreteness; the even case follows analogously since this only introduces

a new Z which equals the previous one)

$$\begin{aligned}
L(\rho, Z) &= \mathrm{tr}(\rho_{AM,n}\Pi_A^{(1)}) \\
&\quad + \mathrm{tr}\left[Z_{A,0}\otimes\mathbb{I}_M\left(\rho_{AM,0} - |\psi_{A,0}\rangle\langle\psi_{A,0}|\otimes\mathbb{I}_M\right)\right] \\
&\quad + \sum_{i\in\{1,3,\ldots n\}} \mathrm{tr}\left[Z_{A,i}\otimes\mathbb{I}_M(\rho_{AM,i} - E_{A,i}U_{A,i}\rho_{AM,i-1}U_{A,i}^\dagger E_{A,i})\right] \\
&\quad + \sum_{i\in\{2,4\ldots n-1\}} \mathrm{tr}\left[Z_{A,i}\otimes\mathbb{I}_M(\rho_{AM,i} - \rho_{AM,i-1})\right] \\
&= \mathrm{tr}(Z_{A,0}|\psi_{A,0}\rangle\langle\psi_{A,0}|) \\
&\quad + \mathrm{tr}\left[\left(Z_{A,0}\otimes\mathbb{I}_M\right)\rho_{AM,0} - \left(Z_{A,1}\otimes\mathbb{I}_M\right)E_{A,1}U_{A,1}\rho_{AM,0}U_{A,1}^\dagger E_{A,1})\right] \\
&\quad + \sum_{i\in\{1,3\ldots n-2\}} \mathrm{tr}\left[\rho_{AM,i}(Z_{A,i} - Z_{A,i+1})\otimes\mathbb{I}_M\right] \\
&\quad + \sum_{i\in\{3,5\ldots n-2\}} \mathrm{tr}\left[-\left(Z_{A,i}\otimes\mathbb{I}_M\right)E_{A,i}U_{A,i}\rho_{AM,i-1}U_{A,i}^\dagger E_{A,i} + \left(Z_{A,i-1}\otimes\mathbb{I}_M\right)\rho_{AM,i-1}\right] \\
&\quad + \mathrm{tr}\left[\left(Z_{A,n}\otimes\mathbb{I}_M\right)\rho_{AM,n} - \rho_{AM,n}\left(\Pi_A^{(1)}\otimes\mathbb{I}_M\right)\right]
\end{aligned}$$

which, after pulling out the ρ terms, can be expressed as

$$\begin{aligned}
L(\rho, Z) &= -\mathrm{tr}(Z_{A,0}|\psi_{A,0}\rangle\langle\psi_{A,0}|) \\
&\quad + \mathrm{tr}\left[\rho_{AM,0}\underbrace{\left(Z_{A,0}\otimes\mathbb{I}_M - U_{A,1}^\dagger E_{A,1}\left(Z_{A,1}\otimes\mathbb{I}_M\right)E_{A,1}U_{A,1}\right)}_{\mathrm{I}}\right] \\
&\quad + \sum_{i\in\{1,3\ldots n-2\}} \mathrm{tr}\left[\rho_{AM,i}\underbrace{(Z_{A,i} - Z_{A,i+1})\otimes\mathbb{I}_M}_{\mathrm{II}}\right] \\
&\quad + \sum_{i\in\{3,5\ldots n-2\}} \left[\mathrm{tr}\rho_{AM,i-1}\underbrace{\left(\left(Z_{A,i-1}\otimes\mathbb{I}_M\right) - U_{A,i}^\dagger E_{A,i}\left(Z_{A,i}\otimes\mathbb{I}_M\right)E_{A,i}U_{A,i}\right)}_{\mathrm{III}}\right] \\
&\quad + \mathrm{tr}\left[\rho_{AM,n}\underbrace{\left(Z_{A,n} - \Pi_A^{(1)}\right)\otimes\mathbb{I}_M}_{\mathrm{IV}}\right].
\end{aligned}$$

Since we maximize L over all $\rho_{AM,i} \geq 0$, to obtain a non-trivial value for L, we must have I, II, III and IV ≤ 0. This yields

$$\begin{aligned}
Z_{A,i-1}\otimes\mathbb{I}_M &\leq U_{A,i}^\dagger E_{A,i}\left(Z_{A,i}\otimes\mathbb{I}_M\right)E_{A,i}U_{A,i} && i\in\{1,3,\ldots n-2\} \\
Z_{A,i} &= Z_{A,i+1} && i\in\{1,3,\ldots n-2\} \\
Z_{A,n} &\leq \Pi_A^{(1)}
\end{aligned}$$

where we used equality in the second equation because, recall, the corresponding primal variables were redundant. We can therefore write $P_B^* \leq \min_{\{Z_{A,i}\}_{i=1}^n} -\mathrm{tr}(Z_{A,0}|\psi_{A,0}\rangle\langle\psi_{A,0}|)$ subject to the aforementioned constraints. Substituting $Z_{A,i}$ with $-Z_{A,i}$ we obtain the desired form. The positivity constraint on $Z_{A,i}$ follows from two observations. First, $Z_{A,n} \geq \Pi_A^{(1)} \geq 0$ as $\Pi_A^{(1)}$ is a POVM element. Second, suppose Z_is are hermitian and U_is are unitary. Then $Z_{i-1} \geq 0$ if $Z_{i-1} \geq U_i Z_i U_i^\dagger$ and $Z_i \geq 0$. The analysis carries over to the case of P_A^*.

The 5th constraint does not change the value of the programme but facilitates further discussions. This is proved next. □

Proof of Proposition 23. To prove the proposition, assume that we are given an optimal solution $\{Z_{A,i}\}_{i=1}^{n}$ to the dual problem (see Theorem 20) satisfying the first four constraints. This, in particular, entails that $P_B^* = \langle\psi_{A,0}| Z_{A,0} |\psi_{A,0}\rangle$. If $|\psi_{A,0}\rangle$ were an eigenket of $Z_{A,0}$ then the claim automatically holds. Let us assume the contrary. If we can construct another operator $Z'_{A,0}$ such that (1) $Z'_{A,0} \geq Z_{A,0}$ and (2) $Z'_{A,0} |\psi_{A,0}\rangle = (P_B^* + \epsilon) |\psi_{A,0}\rangle$ where ϵ can be made arbitrarily small, then we would have proven the proposition. This is because we could use the matrices $\{Z'_{A,0}, Z_{A,1}, \dots Z_{A,n}\}$ which yield the same bound as $\epsilon \to 0$.

We drop the subscript $A, 0$ in the remaining proof, viz. $Z_{A,0} \to Z$, $Z'_{A,0} \to Z'$ and $|\psi_{A,0}\rangle \to |\psi\rangle$. We show that $Z' := \underbrace{(\langle\psi| Z |\psi\rangle + \epsilon)}_{:=\beta} |\psi\rangle\langle\psi| + \Lambda (\mathbb{I} - |\psi\rangle\langle\psi|)$ satisfies the requirements, where Λ is shown to be finite for every $\epsilon > 0$. Z' satisfies requirement (2) by construction. We impose the requirement (1) by enforcing, for all normalised vectors $|\phi\rangle \in \mathcal{A}$, that

$$\begin{aligned}\langle\phi| (Z' - Z) |\phi\rangle &= (P_B^* + \epsilon) |\langle\phi|\psi\rangle|^2 + \Lambda \left(1 - |\langle\phi|\psi\rangle|^2\right) \\ &\quad - \langle\phi| Z |\phi\rangle \geq 0.\end{aligned}$$

Writing $|\phi\rangle = a |\psi\rangle + \bar{a} \left|\psi^\perp\right\rangle$ where $|a|^2 + \bar{a}^2 = 1$ and $\langle\psi|\psi^\perp\rangle = 0$ we have

$$\langle\phi| Z |\phi\rangle = |a|^2 \cancelto{P_B^*}{\langle\psi| Z |\psi\rangle} + \bar{a}^2 \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle + \bar{a} \left(a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.}\right).$$

We restricted to $\bar{a} \in \mathbb{R}$ as the phase can be absorbed in $\left|\psi^\perp\right\rangle$. Substituting this, we obtain

$$\begin{aligned}\langle\phi| (Z' - Z) |\phi\rangle &= (P_B^* + \epsilon) |a|^2 + \Lambda \cancelto{\bar{a}^2}{(1 - |a|^2)} - |a|^2 P_B^* \\ &\quad - \bar{a}^2 \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle - \bar{a} \left(a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.}\right) \\ &= |a|^2 \epsilon + \bar{a}^2 \left(\Lambda - \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle\right) - \bar{a} \left(a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.}\right) \geq 0.\end{aligned}$$

If $a = 0$, then we can simply pick any $\Lambda > \|Z\|$. If $a \neq 0$, we view the equation as a quadratic equation in $\bar{a}$. Since we already require $\Lambda > \|Z\|$, the coefficient of $\bar{a}^2$ is taken to be positive. If we can ensure that the quadratic has no roots then the inequality is guaranteed to hold. To this end, we require that the discriminant is negative. This yields

$$\begin{aligned}&(a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.})^2 - 4 |a|^2 \epsilon (\Lambda - \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle) \leq 0 \\ \iff \quad &\frac{1}{4 |a|^2 \epsilon} (a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.})^2 + \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle \leq \Lambda.\end{aligned}$$

Using $x + \text{h.c.} = 2\Re(x) \leq 2 \|x\|$ for any complex number x, we have

$$\frac{1}{4 |a|^2 \epsilon} (a \langle\psi| Z \left|\psi^\perp\right\rangle + \text{h.c.})^2 \leq \frac{\cancel{4 |a|^2}}{\cancel{4 |a|^2} \epsilon} \|Z\|$$

which means that it suffices to have

$$\frac{\|Z\|}{\epsilon} \leq \frac{\|Z\|}{\epsilon} + \left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle \leq \Lambda$$

to ensure the discriminant is negative. Using $\left\langle\psi^\perp\right| Z \left|\psi^\perp\right\rangle \leq \|Z\|$, we conclude that it suffices to set $\Lambda = \left(\frac{1}{\epsilon} + 1\right) \|Z\|$.

One can proceed analogously for the P_A^* case. □

§ A.2 (Time Dependent) Point Games with EBM transitions/functions

Proof of Proposition 29. Assume we are given a WCF protocol in the standard form, together with the certificates $\{Z_{A,i}\}_{i=1}^n$ and $\{Z_{B,i}\}_{i=1}^n$ (see Theorem 20) witnessing the bias P_A^* and P_B^*. Using Proposition 23 we can use $Z'_{A,0}$ and $Z'_{B,0}$ instead of $Z_{A,0}$ and $Z_{B,0}$ which admit $|\psi_{A,0}\rangle$ and $|\psi_{B,0}\rangle$ as eigenkets with eigenvalues $\beta = P_B^* + \delta$ and $\alpha = P_A^* + \delta$, respectively.

We use the reversed time convention[1]: $Z_A^{(i)} := Z_{A,n-i}$, $Z_B^{(i)} := Z_{B,n-i}$ and similarly $\left|\psi^{(i)}\right\rangle := |\psi_{n-i}\rangle$. We begin with establishing the boundary conditions of the EBM point game. We define $p_0 := \text{Prob}[Z_A^{(0)}, Z_B^{(0)}, |\psi^{(0)}\rangle]$. Consider $Z_A^{(0)} \otimes \mathbb{I}_M \otimes Z_B^{(0)}$ along with the state $\left|\psi^{(0)}\right\rangle$. Recall (see Theorem 20) that $Z_A^{(0)} = \Pi_A^{(1)}$ (Alice's POVM element corresponding to 'Bob wins') and $Z_B^{(0)} = \Pi_B^{(0)}$ (Bob's POVM element corresponding to 'Alice wins'). A WCF protocol by definition (see Definition 17) satisfies

$$\text{tr}(\Pi_A^{(1)} \otimes \mathbb{I}_M \otimes \Pi_B^{(1)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) = \text{tr}(\Pi_A^{(0)} \otimes \mathbb{I}_M \otimes \Pi_B^{(0)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) = \frac{1}{2}$$

and

$$\text{tr}(\Pi_A^{(1)} \otimes \mathbb{I}_M \otimes \Pi_B^{(0)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) = \text{tr}(\Pi_A^{(1)} \otimes \mathbb{I}_M \otimes \Pi_B^{(0)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) = 0$$

corresponding to perfect agreement and equal probabilities of the players winning. Thus,

$$\begin{aligned} p_0 &= \text{Prob}[1.\Pi_A^{(1)} + 0.\underbrace{(\mathbb{I} - \Pi_A^{(1)})}_{:=\Pi_A^{(0)}}, 1.\Pi_B^{(0)} + 0.\underbrace{(\mathbb{I} - \Pi_B^{(0)})}_{:=\Pi_B^{(1)}}, \left|\psi^{(0)}\right\rangle] \\ &= \cancel{\text{tr}(\Pi_A^{(1)} \otimes \mathbb{I}_M \otimes \Pi_B^{(0)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) [\![1,1]\!]} + \cancel{\text{tr}(\Pi_A^{(0)} \otimes \mathbb{I}_M \otimes \Pi_B^{(1)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) [\![0,0]\!]} + \\ &\quad \text{tr}(\Pi_A^{(1)} \otimes \mathbb{I}_M \otimes \Pi_B^{(1)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) [\![1,0]\!] + \text{tr}(\Pi_A^{(0)} \otimes \mathbb{I}_M \otimes \Pi_B^{(0)} \left|\psi^{(0)}\right\rangle \left\langle\psi^{(0)}\right|) [\![0,1]\!] \\ &= \frac{1}{2} [\![1,0]\!] + \frac{1}{2} [\![0,1]\!] . \end{aligned}$$

Analogously, we define $p_n := \text{Prob}[Z_A^{(n)}, Z_B^{(n)}, |\psi^{(n)}\rangle]$. Consider $Z_A^{(n)} \otimes \mathbb{I}_M \otimes Z_B^{(n)}$ and $\left|\psi^{(n)}\right\rangle$. Recall that $\left|\psi^{(n)}\right\rangle = |\psi_{A,0}\rangle \otimes |\psi_{M,0}\rangle \otimes |\psi_{B,0}\rangle$ (see Definition 17) and as stated at the outset, we can take $Z_A^{(n)} = \beta |\psi_{A,0}\rangle \langle\psi_{A,0}| + \Lambda(\mathbb{I} - |\psi_{A,0}\rangle \langle\psi_{A,0}|)$, $Z_B^{(n)} = \alpha |\psi_{B,0}\rangle \langle\psi_{B,0}| + \Lambda(\mathbb{I} - |\psi_{B,0}\rangle \langle\psi_{B,0}|)$. Noting that $\left|\psi_{A/B,0}\right\rangle$ are normalised vectors, it follows that $p_n = 1. [\![\beta, \alpha]\!]$.

Finally, we define

$$p_i := \text{Prob}[Z_A^{(i)}, Z_B^{(i)}, \left|\psi^{(i)}\right\rangle] \tag{A.1}$$

and show that $p_{i-1} \to p_i$ is an EBM transition (either horizontal or vertical, depending on i). Suppose i is such that $Z_B^{(i-1)} = Z_B^{(i)}$ while

$$Z_A^{(i-1)} \leq U_A^{(i)\dagger} E_A^{(i)} Z_A^{(i)} E_A^{(i)} U_A^{(i)} \tag{A.2}$$

(this depends on whether n is odd or even; in the other case, $Z_A^{(i-1)} = Z_A^{(i)}$ while $Z_B^{(i-1)}$ will satisfy an inequality). We also have $\left|\psi^{(i-1)}\right\rangle = U^{(i)\dagger} \left|\psi^{(i)}\right\rangle$. Then, note that

$$\text{Prob}[U_A^{(i)\dagger} E_A^{(i)} Z_A^{(i)} E_A^{(i)} U_A^{(i)}, Z_B^{(i)}, U_A^{(i)\dagger} \left|\psi^{(i)}\right\rangle] = \text{Prob}[Z_A^{(i)}, Z_B^{(i)}, \left|\psi^{(i)}\right\rangle]. \tag{A.3}$$

[1] notation conflict: we have used $\Pi_A^{(0/1)}$ to denote the POVM elements of Alice's final measurement; not to be confused with the reversed time convention which also uses bracketed indices like (i).

To see this, consider

$$Z = \sum_{z \in \mathrm{sp}(Z)} z \Pi^{[z]}$$

where $\Pi^{[z]}$ is the projector onto the z-eigenvalued subspace of Z and note that

$$\mathrm{Prob}[Z, |\psi\rangle] = \sum_z z \langle\psi| \Pi^{[z]} |\psi\rangle = \sum_z z \langle\psi| UU^\dagger \Pi^{[z]} UU^\dagger |\psi\rangle = \mathrm{Prob}[U^\dagger Z U, U|\psi\rangle].$$

Noting that the projector E leaves $|\psi\rangle$ unchanged and so $\mathrm{Prob}[Z, |\psi\rangle] = \mathrm{Prob}[EZE, |\psi\rangle]$, the result easily extends to the bi-variate Prob establishing Equation (A.3). The equation is useful because once combined with Equation (A.1), we can write $p_i = \mathrm{Prob}[U_A^{(i)\dagger} E_A^{(i)} Z_A^{(i)} E_A^{(i)} U_A^{(i)}, Z_B^{(i-1)}, |\psi^{(i-1)}\rangle]$ while $p_{i-1} = \mathrm{Prob}[Z_A^{(i-1)}, Z_B^{(i-1)}, |\psi^{(i-1)}\rangle]$. For the transition from $p_{i-1} \to p_i$ to be EBM, we now only need to show that Equation (A.2) holds which we are given to be true. □

§ A.3 Time Independent Point Games (TIPGs)

Proof of Theorem 39. We will need the following.

Lemma 168. *If the transition $p' \to q'$ is transitively valid and $\zeta : \mathbb{R}_{\geq} \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ is a non-negative function with finite support, then the transition $\delta \cdot p' + \zeta \to \delta \cdot q' + \zeta$ is also a transitively valid for all $\delta > 0$.*

Proof. It is enough to show that the statement holds for valid line transitions (simply because a valid transition is a valid line transition along one of the coordinates). Suppose that $l \to r$ is a valid line transition. Let $\xi : \mathbb{R}_{\geq} \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ be a finitely supported function. Then, $\delta \cdot l + \xi \to \delta \cdot r + \xi$ is valid because $\delta \cdot (r - l)$ is given to be a valid function. □

We prove the theorem in three parts.

Part 1 | Aim: To show that a transition from $\frac{1}{2}[\![0,1]\!] + \frac{1}{2}[\![1,0]\!] \to [\![\beta, \alpha]\!]$ is transitively valid in the presence of an extra set of points (together called a catalyst), given the horizontally valid and vertically valid functions a and b respectively. How these extra points can be created and destroyed are the subjects of part two and three.

We first show that this set of extra points can be taken to be b^-, i.e. we show that each transition in the following is valid:

$$\frac{1}{2}([\![0,1]\!] + [\![1,0]\!]) + b^- \to \frac{1}{2}([\![0,1]\!] + [\![1,0]\!]) + b^+ \to [\![\beta, \alpha]\!] + b^-. \tag{A.4}$$

The first transition is valid because $b^- \to b^+$ is given to be valid (as b is a valid function) and using Lemma 168. Showing that the second transition is valid takes some work (and uses the fact that $a^- \to a^+$ is valid). Note that one can rewrite $a + b = -\frac{1}{2}([\![0,1]\!] + [\![1,0]\!]) + [\![\beta, \alpha]\!]$ as

$$\frac{1}{2}([\![0,1]\!] + [\![1,0]\!]) + b^+ = \underbrace{[\![\beta, \alpha]\!] - a^+ + b^-}_{:=\zeta} + a^-. \tag{A.5}$$

While not apparent at first, ζ is in fact a non-negative function. We return to this in a moment. Note that $\zeta + a^- \to \zeta + a^+$ is valid which we can (using the Equation (A.5) for the LHS (of the second transition in Equation (A.4)) and definition of ζ for the RHS) rewrite as

$$\frac{1}{2}\left([\![0,1]\!] + [\![1,0]\!]\right) + b^+ \to [\![\beta,\alpha]\!] \cancel{-a^+} + b^- \cancel{+a^+}$$

which establishes that the second transition is also valid, granted we can show $\zeta \geq 0$. To see this, note that the negativity could only be present at the points in $\text{supp}(a^+)$ and therefore if we add to it a non-negative function with no support there, such as a^-, we have $\zeta \geq 0 \iff \zeta + a^- \geq 0$. Using Equation (A.5) we see that $\zeta + a^- \geq 0$ completing this part of the demonstration.

Next, we show that we can scale down the weight of b^- to $\delta \cdot b^-$ and still have the transition stay *transitively* valid.

Lemma 169. *Consider a transitively valid transition $p+\xi \to q+\xi$ where $\xi \geq 0$. Then for any/all $\gamma > 0$, the transition $p + \gamma\xi \to q + \gamma\xi$ is transitively valid.*

Proof. Let γ' be the largest inverse of an integer satisfying $\gamma \geq \gamma'$ (this ensures we use the fewest number of steps; should become clear shortly), i.e. $\gamma' = 1/\lceil 1/\gamma \rceil$. The key observation here is that the following transition is transitively valid:

$$p + \gamma'\xi \to (1-\gamma')p + \gamma' q + \gamma'\xi.$$

To see this, note that

$$p + \gamma'\xi = (1-\gamma')p + \gamma'(p+\xi) \to (1-\gamma')p + \gamma'(q+\xi)$$

is transitively valid by using Lemma 168, the fact that $(1-\gamma')p \geq 0$ and that $p + \xi \to q + \xi$ is given to be transitively valid. We can repeat this process to write

$$\begin{aligned} p + \gamma'\xi &\to (1-\gamma')p + \gamma' q + \gamma'\xi \\ &\to (1-2\gamma')p + \gamma' q + \gamma' q + \gamma'\xi = (1-2\gamma')p + 2\gamma' q + \gamma'\xi \\ &\to (1-3\gamma')p + 3\gamma' q + \gamma'\xi \\ &\vdots \quad \text{after } 1/\gamma' \text{ steps} \\ &\to q + \gamma'\xi. \end{aligned}$$

We are almost there. To complete the argument we simply add the residual weight on ξ, i.e. $(\gamma - \gamma')\,\xi$ to both sides (allowed by Lemma 168). □

We substitute ξ with b^- in Lemma 169 to obtain that for any $\gamma > 0$

$$\frac{1}{2}\left([\![0,1]\!] + [\![1,0]\!]\right) + \gamma b^+ \to [\![\beta,\alpha]\!] + \gamma b^- \tag{A.6}$$

is a transitively valid move with at most $2\lceil 1/\gamma \rceil$ intermediate transitions (two for each step; see the proof of Lemma 169). This completes the first part of the proof.

Part 2 | Aim: Construct the small weighted catalyst, γb^-.

Let

$$m = \min_{(x,y)\in \text{supp}(b^-)} \{\max\{x,y\}\}$$

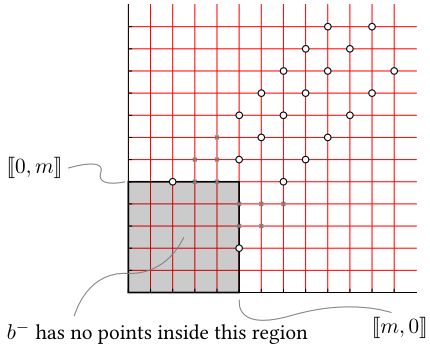


Figure A.1: Significance of m in the proof of equivalence of TIPG and TDPG.

which essentially means that no point in the support of b^- lies in the shaded region of the graph in Figure A.1. It is clear then that given a point in b^-, it can either be reached by raising $[\![0, m]\!]$ or by raising $[\![m, 0]\!]$. More formally, for some $c, d \in \mathbb{R}_{\geq 0}$ satisfying $c + d = \|b^-\|$, the transition

$$c\,[\![0, m]\!] + d\,[\![m, 0]\!] \to b^- \tag{A.7}$$

is transitively valid (by two valid transitions) where the norm we use here is the one-norm, i.e. $\|b^-\| = \sum_{(x,y)\in \text{supp}(b^-)} |b(x, y)|$. Our task therefore is to show that $\frac{1}{2}\,([\![0, 1]\!] + [\![1, 0]\!])$ to $\gamma'(c\,[\![0, m]\!] + d\,[\![m, 0]\!])$ is transitively valid for some $\gamma' > 0$ (normalisation). If $m > 1$, then we can simply raise some weight from $[\![0, 1]\!]$ and $[\![1, 0]\!]$. Otherwise if $m < 1$ then we can split, i.e. note that for appropriately large $m_x, m_y \in \mathbb{R}_{\geq 0}$, the following transitions are valid splits:

$$[\![0, 1]\!] \to \frac{cm}{c+d}\,[\![0, m]\!] + \frac{d + c(1-m)}{c+d}\,[\![0, m_y]\!] \tag{A.8}$$

$$[\![1, 0]\!] \to \frac{dm}{c+d}\,[\![m, 0]\!] + \frac{c + d(1-m)}{c+d}\,[\![m_x, 0]\!] \tag{A.9}$$

assuming $c, d > 0$. Note that we must have $m > 0$ else we can't use a split to put weight on $[\![0, 0]\!]$. However, no TIPG can involve a $[\![0, 0]\!]$ point (see Lemma 3.23 of [3]). To see this, observe that the weight is conserved. Further, it suffices to show that $1 \geq \frac{cm}{c+d}\frac{1}{m} + \frac{d+c(1-m)}{c+d}\frac{1}{m_y}$ for some m_y and clearly, as $m_y \to \infty$, the inequality becomes $1 \geq \frac{c}{c+d}$ which strictly holds (and hence there is a finite m_y). We can therefore write

$$\begin{aligned}
\frac{1}{2}\,[\![0, 1]\!] + \frac{1}{2}\,[\![1, 0]\!] \to{}& \frac{1-\delta}{2}\,[\![0, 1]\!] + \frac{\delta}{2}\left(\frac{cm}{c+d}\,[\![0, m]\!] + \frac{d + c(1-m)}{c+d}\,[\![0, m_y]\!]\right) + && \text{using Eq. (A.8)}\\
&\frac{1-\delta}{2}\,[\![1, 0]\!] + \frac{\delta}{2}\left(\frac{dm}{c+d}\,[\![m, 0]\!] + \frac{c + d(1-m)}{c+d}\,[\![m_x, 0]\!]\right) && \text{using Eq. (A.9)}\\
\to{}& (1-\delta)\left(\frac{[\![0, 1]\!]}{2} + \frac{[\![1, 0]\!]}{2} + \frac{\delta \cdot m}{2(1-\delta)(c+d)}b^-\right) + && \text{using Eq. (A.7)}\\
&\frac{\delta}{2}\left(\frac{d + c(1-m)}{c+d}\,[\![0, m_y]\!] + \frac{c + d(1-m)}{c+d}\,[\![m_x, 0]\!]\right)\\
\to{}& (1-\delta)\,[\![\beta, \alpha]\!] + \frac{\delta \cdot m}{c+d}b^- + && \text{using Eq. (A.6)}\\
&\frac{\delta}{2}\left(\frac{d + c(1-m)}{c+d}\,[\![0, m_y]\!] + \frac{c + d(1-m)}{c+d}\,[\![m_x, 0]\!]\right).
\end{aligned}$$

Let us also keep track of the number of valid transitions we needed in the aforesaid. The first step required two (one for splitting $[\![0, 1]\!]$ and one for $[\![1, 0]\!]$), the second step again required two (a horizontal set of raises and a vertical set of raises to obtain the configuration of b^-), the third required $2\lceil 1/\gamma \rceil$

where γ was the coefficient of b^-. Thus the total number of valid transition we used to show that

$$\frac{1}{2}[\![0,1]\!] + \frac{1}{2}[\![1,0]\!] \to (1-\delta)[\![\beta,\alpha]\!] + \delta\xi \tag{A.10}$$

is transitively valid, is $2+2+2\left\lceil \frac{2(1-\delta)\|b^-\|}{m} \right\rceil$ where we defined $\xi := \frac{m}{2(c+d)}b^- + \frac{d+c(1-m)}{2(c+d)}[\![0,m_y]\!] + \frac{c+d(1-m)}{2(c+d)}[\![m_x,0]\!]$, and used $\|b^-\| = c+d$.

Part 3 | Aim: Absorb the catalyst at a small cost to the bias.

In this last part, we show how to get rid of the $\delta\xi$ part in Equation (A.10) by slightly increasing the bias. We can assume without loss of generality that the state ξ has all the weight on a single point, say $[\![n_x,n_y]\!]$, because if this is not the case, we simply raise all the points by choosing n_x and n_y to be large enough. If we try to directly merge, we must align either along the x-axis or the y-axis by a finite raise, thereby destroying the security for one of the players. Instead (see Figure A.2), we take a small weight (step I) from the point at $[\![\beta,\alpha]\!]$ and align it vertically with $[\![n_x,n_y]\!]$. We merge these (step II) to get $[\![n_x,\alpha+\epsilon]\!]$ by having chosen the small weight appropriately. Since the y-coordinate is now close to α, we raise the remaining weight at $[\![\beta,\alpha]\!]$ to $[\![\beta,\alpha+\epsilon]\!]$ (step III) and then merge with $[\![n_x,\alpha+\epsilon]\!]$ to obtain $[\![\beta+\epsilon,\alpha+\epsilon]\!]$ (step IV). The procedure is formalised by the following lemma. We now state and prove this result formally.

Lemma 170. *Given any $\beta,\alpha,\epsilon > 0$, a finitely supported function $\xi : \mathbb{R}_{\geq} \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ satisfying $\|\xi\| = 1$, there exists a δ satisfying $0 < \delta < 1$ such that*

$$(1-\delta)[\![\beta,\alpha]\!] + \delta\xi \to [\![\beta+\epsilon,\alpha+\epsilon]\!]$$

is transitively valid.

Proof. We raise all the points in ξ to the point $[\![n_x,n_y]\!]$. We define δ and δ' to be such that the following two transitions (merges) are valid:

$$\delta'[\![n_x,\alpha]\!] + \delta[\![n_x,n_y]\!] \to (\delta+\delta')[\![n_x,\alpha+\epsilon]\!], \tag{A.11}$$

$$(1-\delta-\delta')[\![\beta,\alpha+\epsilon]\!] + (\delta'+\delta)[\![n_x,\alpha+\epsilon]\!] \to [\![\beta+\epsilon,\alpha+\epsilon]\!]. \tag{A.12}$$

The (saturated) merge conditions are

$$\begin{aligned} \delta'\alpha + \delta n_y &= (\delta+\delta')(\alpha+\epsilon) \\ (1-\delta-\delta')\beta + (\delta'+\delta)n_x &= \beta+\epsilon \end{aligned}$$

using which one can solve for δ and δ' as

$$\begin{aligned} \delta &= \frac{\epsilon^2}{(n_x-\beta)(n_y-\alpha)} \\ \delta' &= \frac{\epsilon}{n_x-\beta}\left(1-\frac{\epsilon}{n_y-\alpha}\right). \end{aligned} \tag{A.13}$$

These results can be combined to show that the following transitions are valid

$$\begin{aligned} (1-\delta)[\![\beta,\alpha]\!] + \delta\xi &\to (1-\delta)[\![\beta,\alpha]\!] + \delta[\![n_x,n_y]\!] && \text{by raising} \\ &\to (1-\delta-\delta')[\![\beta,\alpha]\!] + \delta'[\![n_x,\alpha]\!] + \delta[\![n_x,n_y]\!] && \text{again, raising} \\ &\to (1-\delta-\delta')[\![\beta,\alpha]\!] + (\delta'+\delta)[\![n_x,\alpha+\epsilon]\!] && \text{by Eq. (A.11)} \\ &\to (1-\delta-\delta')[\![\beta,\alpha+\epsilon]\!] + (\delta'+\delta)[\![n_x,\alpha+\epsilon]\!] && \text{by raising} \\ &\to [\![\beta+\epsilon,\alpha+\epsilon]\!] && \text{by Eq. (A.12).} \end{aligned}$$

This entails $(1-\delta)[\![\beta,\alpha]\!] \to [\![\beta+\epsilon,\alpha+\epsilon]\!]$ is transitively valid as asserted. Further, it is composed of at most $2+1+1+1+1 = 6$ valid transitions. □

Combining Lemma 170 with Equation (A.6) we conclude that

$$\frac{1}{2}[\![0,1]\!]+\frac{1}{2}[\![1,0]\!]\to[\![\beta+\epsilon,\alpha+\epsilon]\!]$$

is transitively valid and is composed of at most

$$10+2\left\lceil\frac{2(1-\delta)\left\|b^-\right\|}{m}\right\rceil \tag{A.14}$$

valid transitions (where δ is related to ϵ and $\xi=b^-$ by Equation (A.13)).

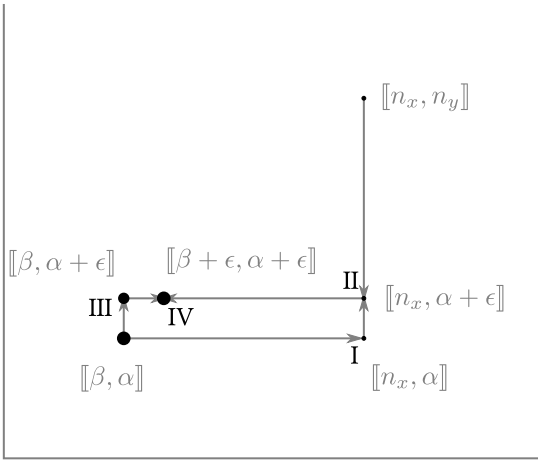


Figure A.2: Absorbing the catalyst at a small cost to the bias.

□

Proof of Corollary 41. We use Equation (A.14). We substitute Γ for n_x and n_y in Equation (A.13), note that $\|b^-\|=\frac{\|b\|}{2}$ (because for a valid function $\|b^-\|=\|b^+\|$) and suppose that $m>1$, to deduce that we need at most $\mathcal{O}\left(\|b\|\,\Gamma^2/\epsilon\right)$ valid transitions. □

§ A.4 Mochon's TIPG achieving bias $\epsilon=1/(4k+2)$

Proof of Lemma 47. It suffices to show that $\sum_{j=\zeta}^{\Gamma}\text{split}(j)=\frac{1}{2}$ and that $\frac{1}{2}>\sum_{j=\zeta}^{\Gamma}\frac{1}{j\omega}\text{split}(j)$ as the former enforces probability conservation and the latter is the split condition (see Example 33). Equating the $\frac{1}{2}$ factor for $\text{split}(j)$ in the inequality we obtain

$$\sum_{j=\zeta}^{\Gamma}\frac{f(0,j\omega)}{\prod_{i=-k}^{k}(j+i)\omega}>\sum_{j=\zeta}^{\Gamma}\frac{f(0,j\omega)}{j\omega\left(\prod_{i=-k}^{k}(j+i)\omega\right)}.$$

Further, we rewrite Equation (2.6) as

$$\begin{aligned}f(0,j\omega)&=(-1)^{k+1}\prod_{i=1}^{k-1}(\alpha-i\omega)(\alpha-i\omega-j\omega)\prod_{i=1}^{k}(\Gamma\omega+i\omega)(\Gamma+i\omega-j\omega)\\&=\left[\prod_{i=1}^{k-1}(j\omega-(\alpha-i\omega))\prod_{i=1}^{k}(\Gamma\omega+i\omega-j\omega)\right]\left[\prod_{i=1}^{k-1}(\alpha-i\omega)\prod_{i=1}^{k}(\Gamma\omega+i\omega)\right].\end{aligned}$$

Substituting it in the inequality (and dividing both sides by the square of the second term in the bracket), we obtain

$$\sum_{j=\zeta}^{\Gamma} p(j\omega) > \sum_{j=\zeta}^{\Gamma} \frac{p(j\omega)}{j\omega}$$

where

$$p(j\omega) := \prod_{i=1}^{k-1} \frac{(j\omega - (\alpha - i\omega))}{(\alpha - i\omega)} \prod_{i=1}^{k} \frac{\Gamma\omega + i\omega - j\omega}{\Gamma\omega + i\omega} \prod_{i=-k}^{k} \frac{1}{j\omega + i\omega}. \tag{A.15}$$

Using Lemma 172 (which we prove shortly), it follows that there exists a choice of ω and Γ such that the inequality is satisfied while $\alpha = \frac{1}{2} + \frac{C}{k}$. □

As was done for the rough calculation, we can simplify the inequality by approximating (1) p by the function $f(x) = \left(\frac{x-\alpha}{\alpha}\right)^{k-1} x^{-2k-1}$ and (2) the sum by an integral.

Lemma 171. *Fix any value for the parameters ω, Γ, α such that*

$$\omega < 1, \quad \Gamma\omega^2 > 1, \quad \textit{and} \quad \alpha > \frac{1}{2}.$$

Define

$$\begin{aligned}
f(x) &= \left(\frac{x-\alpha}{\alpha}\right)^{k-1} x^{-2k-1} \\
\tilde{f}(x) &= \frac{f(x)}{x} \\
\epsilon_l(\omega, \Gamma) &= \Gamma\omega^4 \left|\frac{\partial^2 f}{\partial x^2}\right|_\infty \\
\epsilon_r(\omega, \Gamma) &= \Gamma\omega^4 \left|\frac{\partial^2 \tilde{f}}{\partial x^2}\right|_\infty \\
E(\omega) &= \frac{(1+2k\omega)^{2k+1}}{(1-4k\omega)^{4k+1}}.
\end{aligned}$$

Then, if (for the choice of the parameters)

$$\int_0^{\Gamma\omega^2} f(\omega)dx - \epsilon_l(\omega, \Gamma) > E(\omega)\left[\int_\alpha^\infty \tilde{f}(x)dx + \epsilon_r(\omega, \Gamma)\right] \tag{A.16}$$

then $\sum_{j=\zeta}^{\Gamma} p(j\omega) > \sum_{j=\zeta}^{\Gamma} \frac{p(j\omega)}{j\omega}$ (for the same choice of the parameters) where p is as defined in Equation (A.15).

We return to the proof of this lemma shortly. We use it to complete the main argument first.

Lemma 172. *For any k, there exists a Γ and an ω such that Equation* (A.16) *is satisfied.*

Proof. We take $\Gamma = \omega^{-3}$ so that as $\omega \to 0$, we have $\Gamma \to \infty$, $\Gamma\omega^2 \to \infty$ and $\Gamma\omega^4 \to \infty$. We have

$$\lim_{\omega\to 0} \int_\alpha^{\Gamma\omega^2} f(x)dx - \epsilon_l(\omega, \Gamma) = \int_\alpha^\infty f(x)dx$$

and

$$\lim_{\omega\to 0} E(\omega)\left[\int_\alpha^\infty \tilde{f}(x)dx + \epsilon_r(\omega,\Gamma)\right] = \int_\alpha^\infty \tilde{f}(x)dx.$$

Using the beta function, we have

$$\int_\alpha^\infty f(x)dx = \frac{B(k+1,k)}{\alpha^{2k+2}} \quad \int_\alpha^\infty \tilde{f}(x)dx = \frac{B(k+2,k)}{\alpha^{2k+3}}.$$

Substituting these in the inequality we had to satisfy, we obtain

$$\int_\alpha^\infty f(x)dx > \int_\alpha^\infty \tilde{f}(x)dx \iff \alpha > \frac{B(k+2,k)}{B(k+1,k)} = \frac{k+1}{2k+1}.$$

We have established that

$$\lim_{\omega\to 0}\int_\alpha^{\Gamma\omega^2} f(x)dx - \epsilon_l(\omega,\Gamma) > \lim_{\omega\to 0} E(\omega)\left[\int_\alpha^\infty \tilde{f}(x)dx + \epsilon_r(\omega,\Gamma)\right],$$

i.e. we can find a small ω and $\Gamma = \omega^{-3}$ such that Equation (A.16) holds for any α satisfying $\alpha > \frac{k+1}{2k+1}$. □

It remains to prove Lemma 171 which we now do.

Proof of Lemma 171. We proceed in two steps. In the **first step**, we approximate p by f. Since $\Gamma\omega^2 > 1$ by assumption, one can conclude that

$$\sum_{j=\zeta}^{\lfloor\Gamma\omega\rfloor} p(j\omega) > \sum_{j=\zeta}^{\lfloor\Gamma\omega\rfloor} \frac{p(j\omega)}{j\omega} \implies \sum_{j=\zeta}^{\Gamma} p(j\omega) > \sum_{j=\zeta}^{\Gamma} \frac{p(j\omega)}{j\omega}.$$

(For notational ease, we suppress the floor function henceforth.) To see this, note that for $j > \Gamma\omega$, we have $j\omega > \Gamma\omega^2 > 1 \implies j\omega > 1$. Therefore, for $j > \Gamma\omega$, we anyway have $p(j\omega) > \frac{p(j\omega)}{j\omega}$ and hence, establishing the sum for j from ζ to $\lfloor\Gamma\omega\rfloor$ is enough.

We show that $\prod_{i=1}^k \frac{\Gamma\omega+i\omega-j\omega}{\Gamma\omega+i\omega}$ is close to 1. We have,

$$1 > \prod_{i=1}^{k} \frac{\Gamma\omega + i\omega - j\omega}{\Gamma\omega + i\omega} > \left(1 - \frac{j}{\Gamma}\right)^k > (1-\omega)^k$$

which trivially follows from noting $\frac{\Gamma\omega+i\omega}{\Gamma\omega+i\omega}^{1} - \frac{j\omega}{\Gamma\omega+i\omega} > \left(1 - \frac{j}{\Gamma}\right)$ and using the premise $j/\Gamma > \omega$. Thus, for all $j \in [\zeta, \Gamma\omega]$ we have

$$(1-\omega)^k \prod_{i=1}^{k-1} \frac{j\omega - (\alpha - i\omega)}{\alpha - i\omega} \prod_{i=-k}^{k} \frac{1}{j\omega + i\omega} < p(j\omega) < \prod_{i=1}^{k-1} \frac{j\omega - (\alpha - i\omega)}{\alpha - i\omega} \prod_{i=-k}^{k} \frac{1}{j\omega + i\omega}$$

where, recall that $p(j\omega)$ is as defined in Equation (A.15) and we used the inequality on its middle term. Picking the smallest term in the product on the left and the largest in the right, we can further write

$$(1-\omega)^k \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega+k\omega}\right)^{2k+1} < p(j\omega) < \left(\frac{j\omega-\alpha+k\omega}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega-k\omega}\right)^{2k+1}.$$

Combining these, we note that to establish

$$\sum_{j=\zeta}^{\Gamma} p(j\omega) \geq \sum_{j=\zeta}^{\Gamma} \frac{p(j\omega)}{j\omega}, \tag{A.17}$$

it suffices to show that the lower bound on $\sum_j p(j\omega)$ is larger than the upper bound on $\sum_j p(j\omega)/j\omega$, i.e.,

$$(1-\omega)^k \sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega+k\omega}\right)^{2k+1} \geq \sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha+k\omega}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega-k\omega}\right)^{2k+1} \frac{1}{j\omega}.$$

We pull out the $k\omega$ terms for reasons that should become clear shortly. We construct an LHS' such that LHS>LHS' and an RHS' such that RHS'>RHS so that given LHS'>RHS' entails LHS>RHS. We begin with the LHS and note that

$$\frac{1}{j\omega+k\omega} = \frac{1}{j\omega}\cdot\frac{1}{\left(1+\frac{k\omega}{j\omega}\right)} > \frac{1}{j\omega}\cdot\frac{1}{1+2k\omega}$$

where we used $j\omega \geq \alpha > \frac{1}{2}$ in the inequality. For the RHS, we first shift the sum to remove the $k\omega$ from the denominator. We have

$$\begin{aligned}
\sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha+k\omega}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega-k\omega}\right)^{2k+1} \frac{1}{j\omega} &= \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega-2k\omega}\right)^{2k+1} \frac{1}{j\omega-k\omega} \\
&< \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega-2k\omega}\right)^{2k+2} \\
&< \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha-k\omega}\right)^{k-1} \left(\frac{1}{j\omega}\cdot\frac{1}{1-4k\omega}\right)^{2k+2} \\
&< \frac{1}{(1-4k\omega)^{3k+1}} \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+2}
\end{aligned}$$

where we used $\frac{1}{j\omega-2k\omega} < \frac{1}{j\omega}\cdot\frac{1}{1-4k\omega}$ and $\frac{1}{\alpha-k\omega} < \frac{1}{\alpha}\cdot\frac{1}{1-2k\omega} < \frac{1}{\alpha}\cdot\frac{1}{1-4k\omega}$ in the last two steps. Combining, we conclude that to establish Equation (A.17), it suffices to have

$$\frac{(1-\omega)^k}{(1-2k\omega)^{2k+1}} \sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+1} > \frac{1}{(1-4k\omega)^{3k+1}} \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega+\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+2}.$$

To cancel off some terms, we note that $(1-\omega) > (1-4k\omega)$ to conclude that

$$\sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+1} > E(\omega) \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+2}$$

(with $E(\omega) = (1+2k\omega)^{2k+1}/(1-4k\omega)^{4k+1}$) implies Equation (A.17).

The **second step** involves the approximation of the sums with integrals. Using the so called ''rectangle method'' we have [Ron 2011[2]]:

$$\begin{aligned}
\omega \sum_{j=\zeta}^{\Gamma\omega} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+1} &> \int_{\alpha}^{\Gamma\omega^2} f(x)dx - \frac{\Gamma\omega^2-\alpha}{24}\omega^2 \left|\frac{\partial^2 f}{\partial x^2}\right|_\infty \\
\omega \sum_{j=\zeta+k}^{\Gamma\omega+k} \left(\frac{j\omega-\alpha}{\alpha}\right)^{k-1} \left(\frac{1}{j\omega}\right)^{2k+2} &< \int_{\alpha+k\omega}^{\Gamma\omega^2+k\omega} \tilde{f}(x)dx + \frac{\Gamma\omega^2-\alpha}{24}\omega^2 \left|\frac{\partial^2 f}{\partial x^2}\right|_\infty
\end{aligned}$$

[2] I couldn't find the exact reference; I took it from Aharonov et al's article [3].

where $\tilde{f}(x) = f(x)/x$. The error terms (excluding the sign) are upper bounded by $\epsilon_l(\omega, \Gamma) := \Gamma\omega^4 \left|\partial^2 f/\partial x^2\right|_\infty$ and $\epsilon_r(\omega, \Gamma) := \Gamma\omega^4 \left|\partial^2 \tilde{f}/\partial x^2\right|_\infty$ because $\omega^2 \left(\frac{\Gamma\omega^2 - \alpha}{24}\right) \leq \omega^2 \left(\frac{\Gamma\omega^2}{24}\right) \leq \Gamma\omega^4$. This proves the assertion[3]. □

[3] Again using the argument with LHS>LHS' and RHS'>RHS we conclude that to establish LHS>RHS, it suffices to establish LHS'>RHS'; the primes here refer to the approximation by integrals.

B

Appendix

TEF, Approaching $1/10$ and EMA

§ B.1 TEF functions = Valid functions = closure of EBM functions

Let the set of TEF functions be a set of finitely supported functions $t = h - g$, such that for the associated transition $h \to g$, the conditions in Theorem 10 (or equivalently Theorem 75) can be satisfied for some unitary U. We assumed that h and g are non-negative functions without common support. Then the following lemma holds.

Lemma 173 (TEF = Closure of EBM = valid)**.** *The set of the TEF functions, the set of valid functions (see Definition 68) and the closure of the set of the EBM functions (see Definition 48) are the same.*

Proof sketch. We start by recalling that the set of EBM functions is an open set. From Definition 48 we can see that the matrix H may have eigenvectors which have no support on $|\psi\rangle$. Consequently, one can consider a sequence of EBM functions t_i such that the $\lim_{i\to\infty} t_i = t$ is well-defined, while the associated matrix $\lim_{i\to\infty} H_i$ has a diverging eigenvalue. Such a case arises, for instance, when we have a merge move in the point game. For concreteness, let x_{g_1}, x_{g_2} be the coordinates of two points that are going to be merged into a single point with coordinate $x_h = p_{g_1} x_{g_1} + p_{g_2} x_{g_2}$, and let p_{g_1}, p_{g_2} be their respective probability weights, with $p_{g_1} + p_{g_2} = 1$. Furthermore, let $t_i = [\![x_h + 1/i]\!] - p_{g_1}[\![x_{g_1}]\!] - p_{g_2}[\![x_{g_2}]\!]$. One can verify that for all finite values of i, t_i is EBM, but its limit $t = [\![x_h]\!] - p_{g_1}[\![x_{g_1}]\!] - p_{g_2}[\![x_{g_2}]\!]$ is not EBM (we omit the details for the sake of brevity), thus concluding that the set of EBM functions is open.

To show that the closure of this set is the same as the set of the TEF functions, we need to establish that the limit of any such sequence belongs to the set of TEF functions. This requires a combination of certain results from Chapter 6. In particular, the relationship between the canonical orthogonal form (COF) and the canonical projective form (CPF) permits one to trade the divergence of such a matrix H for appropriate projectors. This is exactly the origin of the projectors E_h that appear in our analysis. The matrices $H \geq G$ and the vector $|\psi\rangle$ corresponding to an EBM transition, can be expressed in the canonical orthogonal form,[1] $X_h \geq O X_g O^T$. Essentially, the same orthogonal matrix O also satisfies the TEF inequality.[2] (Equation (1.9)) The TEF inequality may, in fact, be seen as the limit where H's eigenvalues diverge to infinity. Thus, the limit t of the sequence t_i indeed belongs to the set of TEF functions and this argument readily extends to all relevant sequences.

Finally, in Chapter 3 we saw how the authors of [3] prove that the set of valid functions is the same as the closure of the set of EBM functions. In particular, they start by observing that the set of EBM functions is a convex cone K, and its dual cone K^* is the set of operator monotone functions. The bi-dual K^{**} is the set of valid functions, and the fact that $K^{**} = \mathrm{cl}(K)$ completes the proof. Since we

[1] Recall that X_h and X_g are diagonal matrices containing the eigenvalues of H and G, respectively (in addition to X_h possibly having a large eigenvalue with multiplicities and X_h possibly having zero eigenvalues)

[2] Observe that the TEF inequality is closely related to the canonical projective form.

just showed that the closure of the set of EBM functions is the same as the set of TEF functions, we can also conclude that the set of valid functions is the same as the set of TEF functions. □

§ B.2 Blink $m \to n$ Transition

Recall that the unitary we had described was of the form $U = |w\rangle\langle v| + |v\rangle\langle w| + \sum |v_i\rangle\langle v_i| + \sum |w_i\rangle\langle w_i|$. It is evident that having a scheme for generating these $|v_i\rangle$ and $|w_i\rangle$ will be useful as we explore more complicated transitions. More precisely, we need to complete a set containing one vector into a complete orthonormal basis. Let us do this first and then return to the analysis of a $3 \to 2$ merge.

Completing an Orthonormal Basis

Consider an orthonormal complete set of basis vectors $\{|g_i\rangle\}$, and a vector $|v\rangle = \frac{\sum_i \sqrt{p_i}|g_i\rangle}{\sqrt{\sum_i p_i}}$. We describe a scheme for constructing vectors $|v_i\rangle$ s.t. $\{|v\rangle, \{|v_i\rangle\}\}$ is a complete orthonormal set of basis vectors. Formally, we can do this inductively. Instead, we do this by examples for that makes it intuitive and demonstrates the generalisable argument right away. The first we define to be

$$|v_1\rangle = \frac{\sqrt{p_1}|g_1\rangle - \frac{p_1}{\sqrt{p_2}}|g_2\rangle}{\sqrt{p_1 + \frac{p_1^2}{p_2}}} \left(= \frac{\sqrt{p_1}|g_1\rangle - \sqrt{p_2}|g_2\rangle}{\sqrt{p_1+p_2}}, \text{ the familiar one} \right)$$

which is manifestly normalised and orthogonal to $|v\rangle$, i.e. $\langle v|v_1\rangle = p_1 - p_1 = 0$. The next vector is

$$|v_2\rangle = \frac{\sqrt{p_1}|g_1\rangle + \sqrt{p_2}|g_2\rangle - \frac{(p_1+p_2)}{\sqrt{p_3}}|g_3\rangle}{\sqrt{p_1+p_2+\frac{(p_1+p_2)^2}{p_3}}}$$

which is again manifestly normalised and orthogonal to $|v_1\rangle$ because $\langle v_2|v_1\rangle = \langle v|v_1\rangle$. $\langle v|v_2\rangle = p_1 + p_2 - (p_1+p_2) = 0$. Similarly one can construct the $(k+1)^{\text{th}}$ basis vector as

$$|v_k\rangle = \frac{\sum_{i=1}^k \sqrt{p_k}|g_k\rangle - \frac{\sum_{i=1}^k p_k}{\sqrt{p_{k+1}}}|g_{k+1}\rangle}{N_k}$$

where the $N_k = \sqrt{\sum_{i=1}^k p_k + \frac{(\sum_{i=1}^k p_k)^2}{p_{k+1}}}$ and obtain the full set.

The Analysis

Back to the analysis. Recall that the constraint equation was

$$\underbrace{\sum x_{h_i}|h_{ii}\rangle\langle h_{ii}|}_{\text{I}} + \underbrace{x\mathbb{I}^{\{g_{ii}\}}}_{\text{II}} \geq \underbrace{\sum x_{g_i} U|g_{ii}\rangle\langle g_{ii}|U^\dagger}_{\text{III}}$$

where we have introduced the notation $|h_{ii}\rangle = |h_i h_i\rangle$ in the interest of efficiency. The $g_1, g_2, g_3 \to h_1, h_2$ transition requires us to know

$$U = |v\rangle\langle w| + |w\rangle\langle v| + |v_1\rangle\langle v_1| + |v_2\rangle\langle v_2| + |w_1\rangle\langle w_1|.$$

Using the procedure above we can evaluate the vectors of interest

$$|v\rangle = \frac{\sqrt{p_{g_1}}\,|g_{11}\rangle + \sqrt{p_{g_2}}\,|g_{22}\rangle + \sqrt{p_{g_3}}\,|g_{33}\rangle}{N_g}$$
$$|v_1\rangle = \frac{\sqrt{p_{g_1}}\,|g_{11}\rangle - \frac{p_{g_1}}{\sqrt{p_{g_2}}}\,|g_{22}\rangle}{N_{g_1}}$$
$$|v_2\rangle = \frac{\sqrt{p_{g_1}}\,|g_{11}\rangle + \sqrt{p_{g_2}}\,|g_{22}\rangle - \frac{(p_{g_1}+p_{g_2})}{\sqrt{p_{g_3}}}\,|g_{33}\rangle}{N_{g_2}}$$
$$|w\rangle = \frac{\sqrt{p_{h_1}}\,|h_{11}\rangle + \sqrt{p_{h_2}}\,|h_{22}\rangle}{N_h}$$
$$|w_1\rangle = \frac{\sqrt{p_{h_2}}\,|h_{11}\rangle - \sqrt{p_{h_1}}\,|h_{22}\rangle}{N_h}$$

where N_g, N_{g_1}, N_{g_2}, N_h are normalisations. In fact we want to express the constraints in this basis. To evaluate the first term we use the above to find

$$|h_{11}\rangle = \frac{\sqrt{p_{h_1}}\,|w\rangle + \sqrt{p_{h_2}}\,|w_1\rangle}{N_h}$$
$$|h_{22}\rangle = \frac{\sqrt{p_{h_2}}\,|w\rangle - \sqrt{p_{h_1}}\,|w_1\rangle}{N_h}$$

which leads to

$$\begin{aligned}
\mathrm{I} &= x_{h_1}\,|h_{11}\rangle\langle h_{11}| + x_{h_2}\,|h_{22}\rangle\langle h_{22}| \\
&= \frac{x_{h_1}}{N_h^2}\left[\begin{array}{c|cc} & \langle w| & \langle w_1| \\ \hline |w\rangle & p_{h_1} & \sqrt{p_{h_1}p_{h_2}} \\ |w_1\rangle & \sqrt{p_{h_1}p_{h_2}} & p_{h_2} \end{array}\right] + \frac{x_{h_2}}{N_h^2}\left[\begin{array}{c|cc} & \langle w| & \langle w_1| \\ \hline |w\rangle & p_{h_2} & -\sqrt{p_{h_1}p_{h_2}} \\ |w_1\rangle & -\sqrt{p_{h_1}p_{h_2}} & p_{h_1} \end{array}\right] \\
&= \frac{1}{N_h^2}\left[\begin{array}{c|cc} & \langle w| & \langle w_1| \\ \hline |w\rangle & p_{h_1}x_{h_1} + p_{h_2}x_{h_2} & \sqrt{p_{h_1}p_{h_2}}(x_{h_1} - x_{h_2}) \\ |w_1\rangle & \sqrt{p_{h_1}p_{h_2}}(x_{h_1} - x_{h_2}) & p_{h_2}x_{h_1} + p_{h_1}x_{h_2} \end{array}\right].
\end{aligned}$$

(Remark: We had made a mistake in this term which was causing the matrix to sometimes become negative; after correction, the matrix seems to be positive for Mochon's f-function based construction) Evaluation of II is nearly trivial for identity can be expressed in any basis and that yields

$$\begin{aligned}
\mathrm{II} &= x(|v\rangle\langle v| + |v_1\rangle\langle v_1| + |v_2\rangle\langle v_2|) \\
&= \left[\begin{array}{c|ccc} & \langle v| & \langle v_1| & \langle v_2| \\ \hline |v\rangle & x & & \\ |v_1\rangle & & x & \\ |v_2\rangle & & & x \end{array}\right].
\end{aligned}$$

For the last term

$$\mathrm{III} = \underbrace{x_{g_1}U\,|g_{11}\rangle\langle g_{11}|\,U^\dagger}_{\text{(i)}} + \underbrace{x_{g_2}U\,|g_{22}\rangle\langle g_{22}|\,U^\dagger}_{\text{(ii)}} + \underbrace{x_{g_3}U\,|g_{33}\rangle\langle g_{33}|\,U^\dagger}_{\text{(iii)}}$$

We evaluate

$$U\,|g_{11}\rangle = \frac{\sqrt{p_{g_1}}}{N_g}\,|w\rangle + \frac{\sqrt{p_{g_1}}}{N_{g_1}}\,|v_1\rangle + \frac{\sqrt{p_{g_1}}}{N_{g_2}}\,|v_2\rangle$$
$$U\,|g_{22}\rangle = \frac{\sqrt{p_{g_2}}}{N_g}\,|w\rangle + \frac{\left(-\frac{p_{g_1}}{\sqrt{p_{g_2}}}\right)}{N_{g_1}}\,|v_1\rangle + \frac{\sqrt{p_{g_2}}}{N_{g_2}}\,|v_2\rangle$$
$$U\,|g_{33}\rangle = \frac{\sqrt{p_{g_3}}}{N_g}\,|w\rangle + 0\,|v_1\rangle + \frac{\left(-\frac{p_{g_1}+g_{g_2}}{\sqrt{p_{g_3}}}\right)}{N_{g_2}}\,|v_2\rangle\,.$$

We must now find each sub term, starting with the most regular

$$\text{(i)} = x_{g_1} p_{g_1} \left[\begin{array}{c|ccc} & \langle v_1| & \langle v_2| & \langle w| \\ \hline |v_1\rangle & \frac{1}{N_{g_1}^2} & \frac{1}{N_{g_1} N_{g_2}} & \frac{1}{N_{g_1} N_g} \\ |v_2\rangle & \frac{1}{N_{g_2} N_{g_1}} & \frac{1}{N_{g_2}^2} & \frac{1}{N_{g_2} N_g} \\ |w\rangle & \frac{1}{N_g N_{g_1}} & \frac{1}{N_g N_{g_2}} & \frac{1}{N_g^2} \end{array}\right].$$

For the second term, we re-write $U|g_{22}\rangle = \sqrt{p_{g_2}} \left(\frac{1}{N_g} |w\rangle - \frac{1}{N'_{g_1}} |v_1\rangle + \frac{1}{N_{g_2}} |v_2\rangle \right)$ where we have defined

$$N'_{g_1} = \frac{p_{g_2}}{p_{g_1}} N_{g_1}$$

to obtain

$$(ii) = x_{g_2} p_{g_2} \left[\begin{array}{c|ccc} & \langle v_1| & \langle v_2| & \langle w| \\ \hline |v_1\rangle & \frac{1}{N'^2_{g_1}} & -\frac{1}{N'_{g_1} N_{g_2}} & -\frac{1}{N'_{g_1} N_g} \\ |v_2\rangle & -\frac{1}{N_{g_2} N'_{g_1}} & \frac{1}{N_{g_2}^2} & \frac{1}{N_{g_2} N_g} \\ |w\rangle & -\frac{1}{N_g N'_{g_1}} & \frac{1}{N_g N_{g_2}} & \frac{1}{N_g^2} \end{array}\right]$$

and finally $U|g_{33}\rangle = \sqrt{p_{g_3}} \left(\frac{1}{N_g} |w\rangle + 0|v_1\rangle - \frac{1}{N'_{g_2}} |v_2\rangle \right)$ with

$$N'_{g_2} = \frac{p_{g_3}}{p_{g_1} + p_{g_2}}$$

to get

$$\text{(iii)} = x_{g_3} p_{g_3} \left[\begin{array}{c|ccc} & \langle v_1| & \langle v_2| & \langle w| \\ \hline |v_1\rangle & & & \\ |v_2\rangle & & \frac{1}{N'^2_{g_2}} & -\frac{1}{N'_{g_2} N_g} \\ |w\rangle & & -\frac{1}{N_g N'_{g_2}} & \frac{1}{N_g^2} \end{array}\right].$$

Now we can combine all of these into a single matrix and try to obtain some simpler constraints.

$$M \stackrel{\text{def}}{=} \left[\begin{array}{c|ccccc} & \langle v| & \langle v_1| & \langle v_2| & \langle w| & \langle w_1| \\ \hline |v\rangle & x & & & & \\ |v_1\rangle & & x - \frac{x_{g_1} p_{g_1}}{N_{g_1}^2} - \frac{x_{g_2} p_{g_2}}{N'^2_{g_1}} & -\frac{x_{g_1} p_{g_1}}{N_{g_1} N_{g_2}} + \frac{x_{g_2} p_{g_2}}{N'_{g_1} N_{g_2}} & -\frac{x_{g_1} p_{g_1}}{N_{g_1} N_g} + \frac{x_{g_2} p_{g_2}}{N'_{g_1} N_g} & \\ |v_2\rangle & & -\frac{x_{g_1} p_{g_1}}{N_{g_2} N_{g_1}} + \frac{x_{g_2} p_{g_2}}{N_{g_2} N'_{g_1}} & x - \frac{x_{g_1} p_{g_1}}{N_{g_2}^2} - \frac{x_{g_2} p_{g_2}}{N_{g_2}^2} - \frac{x_{g_3} p_{g_3}}{N'^2_{g_2}} & -\frac{x_{g_1} p_{g_1}}{N_{g_2} N_g} - \frac{x_{g_2} p_{g_2}}{N_{g_2} N_g} + \frac{x_{g_3} p_{g_3}}{N'_{g_2} N_g} & \\ |w\rangle & & -\frac{x_{g_1} p_{g_1}}{N_g N_{g_1}} + \frac{x_{g_2} p_{g_2}}{N_g N'_{g_1}} & -\frac{x_{g_1} p_{g_1}}{N_g N_{g_2}} - \frac{x_{g_2} p_{g_2}}{N_g N_{g_2}} + \frac{x_{g_3} p_{g_3}}{N_g N'_{g_2}} & \frac{p_{h_1} x_{h_1} + p_{h_2} x_{h_2}}{N_h^2} - \frac{1}{N_g^2} \sum_i x_{g_i} p_{g_i} & \frac{\sqrt{p_{h_1} p_{h_2}}}{N_h^2} (x_{h_1} - x_{h_2}) \\ |w_1\rangle & & & & \frac{\sqrt{p_{h_1} p_{h_2}}}{N_h^2} (x_{h_1} - x_{h_2}) & \frac{p_{h_2} x_{h_1} + p_{h_1} x_{h_2}}{N_h^2} \end{array}\right] \geq 0.$$

Despite this appearing to be a complicated expression, we can conclude that it will always be so that larger the x looser will be the constraint. To show this and to simplify this calculation, note that M can be split into a scalar condition, $x \geq 0$ (from the $|v\rangle\langle v|$ part) and a sub-matrix which we choose to write as

$$\left[\begin{array}{c|cc|cc} & \langle v_1| & \langle v_2| & \langle w| & \langle w_1| \\ \hline |v_1\rangle & & & & \\ |v_2\rangle & & C & & B^T \\ \hline |w\rangle & & & & \\ |w_1\rangle & & B & & A \end{array}\right] \geq 0.$$

Now since $\left[\begin{array}{cc} C & B^T \\ B & A \end{array}\right] \geq 0 \iff \left[\begin{array}{cc} A & B \\ B^T & C \end{array}\right] \geq 0 \iff C \geq 0, A - BC^{-1}B^T \geq 0, (\mathbb{I} - CC^{-1})B^T = 0$ using Shur's Complement condition for positivity where C^{-1} is supposed to be the

generalised inverse. Since x is in our hands, we can take it to be sufficiently large so that $C > 0$ and thereby make sure that $\mathbb{I} - CC^{-1} = 0$. Evidently then, the only condition of interest is

$$A - BC^{-1}B^T \geq 0.$$

We can do even better than this actually. Note that if $C > 0$ then $C^{-1} > 0$ and that the second term is of the form

$$\underbrace{\begin{bmatrix} a & b \\ 0 & 0 \end{bmatrix}}_{B}\underbrace{\begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}}_{C^{-1}}\underbrace{\begin{bmatrix} a & 0 \\ b & 0 \end{bmatrix}}_{B^T} = \begin{bmatrix} \begin{bmatrix} a & b \end{bmatrix}\begin{bmatrix} \alpha & \gamma \\ \gamma & \beta \end{bmatrix}\begin{bmatrix} a \\ b \end{bmatrix} & 0 \\ 0 & 0 \end{bmatrix} \geq 0$$

because $C^{-1} > 0$. We can therefore write the constraint equation as

$$A \geq BC^{-1}B^T \geq 0$$

and note that $A \geq 0$ is a necessary condition. This also becomes a sufficient condition in the limit that $x \to \infty$ because $C^{-1} \to 0$ in that case. We have thereby reduced the analysis to simply checking if

$$\begin{bmatrix} \frac{p_{h_1}x_{h_1}+p_{h_2}x_{h_2}}{N_h^2} - \frac{1}{N_g^2}\sum_i x_{g_i}p_{g_i} & \frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}(x_{h_1} - x_{h_2}) \\ \frac{\sqrt{p_{h_1}p_{h_2}}}{N_h^2}(x_{h_1} - x_{h_2}) & \frac{p_{h_2}x_{h_1}+p_{h_1}x_{h_2}}{N_h^2} \end{bmatrix} \geq 0.$$

This being a 2×2 matrix can be checked for positivity by the trace and determinant method. Another possibility is the use of Schur's Complement conditions again. Here, however, we intend to use a more general technique (similar to the one used in the split analysis). Let us introduce

$$\langle x_g \rangle \stackrel{\text{def}}{=} \frac{1}{N_g^2}\sum_i x_{g_i}p_{g_i}, \left\langle \frac{1}{x_h} \right\rangle \stackrel{\text{def}}{=} \frac{1}{N_h^2}\sum_i \frac{p_{h_i}}{x_{h_i}}$$

and recall/note that term (I) and one element from term (III) constitute matrix A, which can also be written as

$$\begin{aligned} A &= x_{h_1}\left|h_{11}\right\rangle\left\langle h_{11}\right| + x_{h_2}\left|h_{22}\right\rangle\left\langle h_{22}\right| - \langle x_g \rangle \left|w\right\rangle\left\langle w\right| \\ &= \begin{array}{c|cc} & \langle h_{11}| & \langle h_{22}| \\ \hline |h_{11}\rangle & x_{h_1} & \\ |h_{22}\rangle & & x_{h_2} \end{array} - \langle x_g \rangle \left|w\right\rangle\left\langle w\right| \end{aligned}$$

Note that this now has the exact same form as that of the split constraint with $x_{g_1} \to \langle x_g \rangle$. We use the same $F - M \geq 0 \iff \mathbb{I} - \sqrt{F}^{-1}M\sqrt{F}^{-1} \geq 0$ for $F > 0$ technique to obtain $\mathbb{I} \geq \langle x_g \rangle \left|w''\right\rangle\left\langle w''\right|$ where $|w''\rangle = \frac{\sqrt{\frac{p_{h_1}}{x_{h_1}}}|h_{11}\rangle + \sqrt{\frac{p_{h_2}}{x_{h_2}}}|h_{22}\rangle}{N_h}$. Normalising this one gets $|w'\rangle = \frac{|w''\rangle}{\sqrt{\left\langle \frac{1}{x_h} \right\rangle}}$ which entails $\mathbb{I} \geq \langle x_g \rangle \left\langle \frac{1}{x_h} \right\rangle \left|w'\right\rangle\left\langle w'\right|$ and that leads us to the final condition

$$\frac{1}{\langle x_g \rangle} \geq \left\langle \frac{1}{x_h} \right\rangle.$$

In fact all the techniques used in reaching this result can be extended to the $m \to n$ transition case as well and so the aforesaid result should hold in general.

§ B.3 Mochon's Assignments

In the following, we assume that $\{x_i\}$ are distinct real numbers.

Lemma (Mochon's Denominator). $\sum_{i=1}^{n} \frac{1}{\prod_{j\neq i}(x_j - x_i)} = 0$ *for* $n \geq 2$.

Proof. We prove this by induction (following Mochon's proof, just optimised for clarity instead of space). For $n = 2$

$$\frac{1}{(x_2 - x_1)} + \frac{1}{(x_1 - x_2)} = 0.$$

Now we show that if the result holds for $n-1$ and it would also hold for n which would complete the inductive proof. We start with noting that

$$\frac{1}{(x_n - x_i)(x_1 - x_i)} = \frac{1}{x_n - x_1}\left[\frac{1}{x_1 - x_i} - \frac{1}{x_n - x_i}\right].$$

This is useful because it helps breaking the product into a sum. My strategy would be to pull off one common term so that we can apply the result to the remaining $n-1$ terms. The expression of interest is

$$\sum_{i=1}^{n} \frac{1}{\prod_{j\neq i}(x_j - x_i)} = \frac{1}{\prod_{j\neq 1}(x_j - x_1)} + \sum_{i=2}^{n-1} \frac{1}{\prod_{j\neq i}(x_j - x_i)} + \frac{1}{\prod_{j\neq n}(x_j - x_n)}$$

where notice that the ith term in the sum (of the second term) can be written as

$$\frac{1}{(x_n - x_i)(x_1 - x_i)\prod_{j\neq i,1,n}(x_j - x_i)} = \frac{1}{x_n - x_1}\left[\frac{1}{\prod_{j\neq i,n}(x_j - x_i)} - \frac{1}{\prod_{j\neq 1,i}(x_j - x_i)}\right].$$

The first term can be written as

$$\frac{1}{(x_n - x_1)\prod_{j\neq 1,n}(x_j - x_1)}$$

while the last can be written as

$$\frac{-1}{(x_n - x_1)\prod_{j\neq n,1}(x_j - x_n)}.$$

Putting all these together, we get

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{1}{\prod_{j\neq i}(x_j - x_i)} \\
&= \frac{1}{(x_n - x_1)}\left[\underbrace{\frac{1}{\prod_{j\neq 1,n}(x_j - x_1)} + \sum_{i=2}^{n-1} \frac{1}{\prod_{j\neq i,n}(x_j - x_i)}} - \underbrace{\sum_{i=2}^{n-1} \frac{1}{\prod_{j\neq 1,i}(x_j - x_i)} - \frac{1}{\prod_{j\neq 1,n}(x_j - x_n)}}\right] \\
&= \frac{1}{(x_n - x_1)}\left[\sum_{i=1}^{n-1} \frac{1}{\prod_{j\neq i,n}(x_j - x_i)} - \sum_{i=2}^{n} \frac{1}{\prod_{j\neq 1,i}(x_j - x_i)}\right]
\end{aligned}$$

where both sums disappear if the result holds for $n-1$. This completes the proof. □

Lemma (Mochon's f-assignment Lemma). $\sum_{i=1}^{n} \frac{f(x_i)}{\prod_{j\neq i}(x_j - x_i)} = 0$ *where* $f(x_i)$ *is of order* $k \leq n-2$.

Proof. Again we do this by induction on k. For $k = 0$ the result holds by the previous result. We assume it holds for order $k-1$ and show using this that it also holds for order k (this proof is also Mochon's). Let $g(x_i)$ be a polynomial of order $k-1$ s.t.

$$\sum_{i=1}^{n} \frac{f(x_i)}{\prod_{j\neq i}(x_j - x_i)} = \sum_{i=1}^{n} \frac{(x_1 - x_i)(x_2 - x_i)\dots(x_k - x_i) - g(x_i)}{\prod_{j\neq i}(x_j - x_i)}.$$

Notice that the first part of the sum disappears for all $1 \leq i \leq k$ because of the numerator. Consequently we can write the aforesaid as

$$\begin{aligned}
&= \sum_{i=k+1}^{n} \frac{(x_1 - x_i)(x_2 - x_i)\dots(x_k - x_i)}{\prod_{j\neq i}(x_j - x_i)} - \sum_{i=1}^{n} \frac{g(x_i)}{\prod_{j\neq i}(x_j - x_i)} \\
&= \sum_{i=k+1}^{n} \frac{1}{\prod_{j\neq i,1,2\dots,k}(x_j - x_i)} \\
&= 0
\end{aligned}$$

where in the first step, the second term becomes zero by assuming the result holds for $k-1$ and in the second step the sum disappears because of the previous result (Mochon's Denominator). Note that $k \leq n-2$ for the aforesaid argument to work because otherwise the last step would become invalid. □

Lemma. $\sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j\neq i}(x_j - x_i)} = (-1)^{n-1}$ *for* $n \geq 2$.

Proof. Let us define $d(n) := \sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j\neq i}(x_j - x_i)}$ to proceed inductively. We can then write

$$d(2) = \frac{x_1}{x_2 - x_1} + \frac{x_2}{x_1 - x_2} = \frac{x_1(x_1 - x_2) + x_2(x_2 - x_1)}{(x_2 - x_1)(x_1 - x_2)} = -1.$$

We assume the result holds for $d(n)$ and write

$$\begin{aligned}
d(n+1) &= \sum_{i=1}^{n+1} \frac{x_i^n}{\prod_{j\neq i}(x_j - x_i)} \\
&= \sum_{i=1}^{n+1} \frac{-(x_{n+1} - x_i)(x_i^{n-1}) + x_{n+1}x_i^{n-1}}{\prod_{j\neq i}(x_j - x_i)} \\
&= -\sum_{i=1}^{n+1} (x_{n+1} - x_i) \frac{x_i^{n-1}}{\prod_{j\neq i}(x_j - x_i)} + x_{n+1} \underbrace{\sum_{i=1}^{n+1} \frac{x_i^{n-1}}{\prod_{j\neq i}(x_j - x_i)}}_{=0 \text{ (Mochon's Denominator)}} \\
&= -\sum_{i=1}^{n} \frac{\cancel{(x_{n+1} - x_i)}}{\cancel{(x_{n+1} - x_i)}} \frac{x_i^{n-1}}{\prod_{j\neq i,n+1}(x_j - x_i)} + \cancel{(x_{n+1} - x_{n+1})} \frac{0}{\prod_{j\neq n+1}(x_j - x_{n+1})} \\
&= -d(n).
\end{aligned}$$

□

Proposition. $\langle x_h \rangle - \langle x_g \rangle = \frac{1}{N_h^2} = \frac{1}{N_g^2}$ *for a Mochon's TDPG assignment with* $k = n-2$ *and coefficient of* x^{n-2} ± 1 *in* $f(x)$. *As above here* $\langle x_h \rangle = \frac{1}{N_h^2} \sum p_{h_i} x_{h_i}$ *and* $\langle x_g \rangle = \frac{1}{N_g^2} \sum p_{g_i} x_{g_i}$.

Proof. Note, to start with, that the coefficient of x^{n-2} being ± 1 is not an artificial requirement because for killing $n-2$ points $f(x)$ will have the form

$$f(x) = (x_{k_1} - x)(x_{k_2} - x) \dots (x_{k_{n-2}} - x) = (-1)^{n-2} x^{n-2} + \tilde{f}(x)$$

where $\tilde{f}$ is a polynomial of order $n-2$. Observe that

$$\begin{aligned} N_h^2 \left(\langle x_h \rangle - \langle x_g \rangle \right) &= \sum_{i=1}^{n} p(x_i) x_i = -\sum_{i=1}^{n} \frac{x_i f(x_i)}{\prod_{j \neq i} (x_j - x_i)} \\ &= -\sum_{i=1}^{n} \frac{x_i (-1)^{n-2} x_i^{n-2}}{\prod_{j \neq i} (x_j - x_i)} - \sum_{i} \frac{\tilde{f}(x_i)}{\prod_{j \neq i} (x_j - x_i)} \\ &= -(-1)^{n-2} \sum_{i=1}^{n} \frac{x_i^{n-1}}{\prod_{j \neq i} (x_j - x_i)} \\ &= 1 \end{aligned}$$

where the second term in the second step vanishes because of Mochon's f-assignment Lemma and the last step follows from the previous result. □

We conclude this section by reproducing Mochon's proof of validity of Mochon's f-assignment.

Proposition 174. *Let* $t = \sum_{i=1}^{n} \frac{-f(x_i)}{\prod_{j \neq i} (x_j - x_i)} [\![x_i]\!]$ *where* f *is a polynomial of degree at most* $n-2$ *satisfying* $f(-\lambda) \geq 0$ *for all* $\lambda \geq 0$ *and* $\{x_1, x_2 \dots x_n\}$ *are distinct non-negative integers. Then* t *is a valid function.*

Proof reproduced from [28]. For t to be valid, we must have $\sum_{x \in \text{supp}(t)} t(x) = 0$ and $\sum_{x \in \text{supp}(t)} \frac{-1}{\lambda + x} t(x) \geq 0$ for all $\lambda \geq 0$. The first condition is equivalent to $\sum_{i=1}^{n} \frac{-f(x_i)}{\prod_{j \neq i} (x_j - x_i)} = 0$ which holds (see Mochon's f-assignment lemma above). The second condition is equivalent to $\sum_{i=1}^{n} \frac{1}{\lambda + x_i} \frac{f(x_i)}{\prod_{j \neq i} (x_j - x_i)} \geq 0$. If we use Mochon's f-assignment lemma with $\{-\lambda, x_1, \dots, x_{2k+1}\}$ then we obtain

$$\begin{aligned} \frac{f(-\lambda)}{\prod_j (x_j - (-\lambda))} + \sum_{i=1}^{n} \frac{f(x_i)}{(-\lambda - x_i) \prod_{j \neq i} (x_j - x_i)} &= 0 \\ \implies \frac{f(-\lambda)}{\prod_j (x_j + \lambda)} &= \sum_{i=1}^{n} \frac{1}{\lambda + x_i} \frac{f(x_i)}{\prod_{j \neq i} (x_j - x_i)} \end{aligned}$$

which in turn is non-negative because $(x_j + \lambda) \geq 0$ as x_j were assumed non-negative and $f(-\lambda) \geq 0$ by assumption for $\lambda \geq 0$. □

C

Appendix

Approaching $1/(4k+2)$

§ C.1 Restricted decomposition into f_0-assignments

The monomial decomposition is not unique. We give another useful decomposition but it only works in a restricted case.

Lemma 175 (f with right-roots to f_0)**.** *Consider a set of real coordinates satisfying* $0 < x_1 < x_2 \cdots < x_n$ *and let* $f(x) = (r_1 - x)(r_2 - x)\dots(r_k - x)$ *where* $k \leq n-2$ *and the roots* $\{r_i\}_{i=1}^k$ *of* f *are right-roots, i.e. they are such that for every root* r_i *there exists a distinct coordinate* $x_j < r_i$*. Let* $t = \sum_{i=1}^n p_i [\![x_i]\!]$ *be the corresponding Mochon's* f*-assignment (justifying the restriction on* k*). Then there exist* f_0*-assignments,* $\{t_{0;j}\}$*, on a subset of* $(x_1, x_2 \dots x_n)$ *such that* $t = \sum_{i=1}^m \alpha_i t_{0;i}$ *where* $\alpha_i > 0$ *is a real number and* $m > 0$ *is an integer.*

Proof. For simplicity, assume that $x_i < r_i$ (for all i) but the argument works in the aforementioned general case. One can then write

$$\begin{aligned} t &= \sum_{i=1}^n \frac{-f(x_i)}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!] \\ &= \sum_{i=1}^n \left(\frac{-(r_1 - x_1)(r_2 - x_i)\dots(r_k - x_i)}{\prod_{j\neq i}(x_j - x_i)} + \frac{-(x_1 - x_i)(r_2 - x_i)\dots(r_k - x_i)}{\prod_{j\neq i}(x_j - x_i)} \right) [\![x_i]\!] \\ &= (r_1 - x_1) \sum_{i=1}^n \frac{-(r_2 - x_i)\dots(r_k - x_i)}{\prod_{j\neq i}(x_j - x_i)} [\![x_i]\!] + \sum_{i=2}^n \frac{-(r_2 - x_i)\dots(r_k - x_i)}{\prod_{j\neq i,1}(x_j - x_i)} [\![x_i]\!], \end{aligned}$$

where the first term has the same form that we started with (except for a positive constant which is irrelevant to the validity condition; see Definition 68) but with the polynomial having one less degree. The second term also has the same form, except that the number of points involved has been reduced. Note how this process relies crucially on the fact that $r_1 - x_1$ is positive (else the term on the left would, by itself, not correspond to a valid move). This process can be repeated until we obtain a sum of f_0 assignments on various subsets of $(x_1, x_2 \dots x_n)$. □

The advantage of this decomposition is that we can immediately apply this result to the f-assignment Mochon uses in the bias $1/10$ game. This is relevant because constructing solutions to f_0-assignments is relatively easy and so they, together with this result, allow us to derive the $1/10$ bias protocol circumventing the perturbative approach introduced in Chapter 4.

Example 176 (The main $1/10$ move.)**.** The key move in Mochon's $1/10$ bias game has its coordinates given by x_0, x_1, x_2, x_3, x_4 and roots given by l_1, r_1, r_2 which satisfy $x_0 < l_1 < x_1 < x_2 < x_3 < x_4 < r_1 < r_2$. Each root is a right root here because $x_0 < l_1$, $x_3 < r_1$, $x_4 < r_2$ for instance.

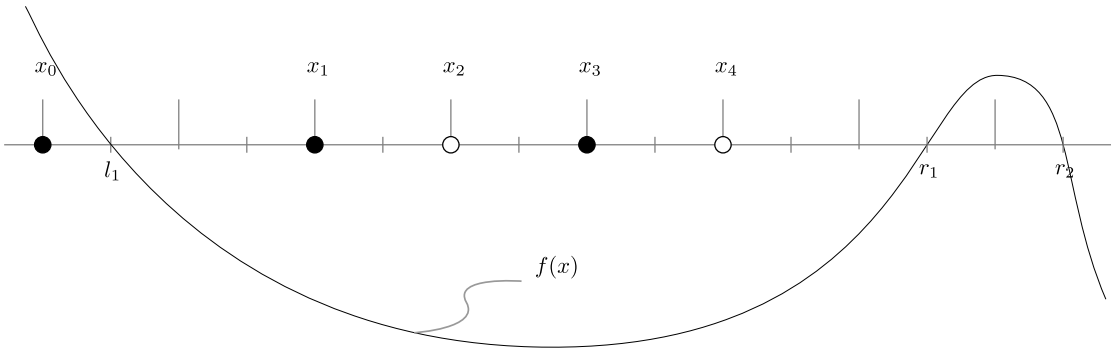


Figure C.1: A typical $1/10$ move involves $n = 5$ points. f has $k = 3$ roots, all of which happen to be left roots. This ceases to be the case for Mochon's games with lower bias.

Hence, this assignment can be expressed as a combination of f_0 assignments defined over subsets of the initial set of coordinates and each f_0 assignment admits a simple solution (see Proposition 165 and Proposition 167).

Another simple example is the class of f-assignments which are merges. We place the roots of f in such a way that all points, except one, have negative weights.

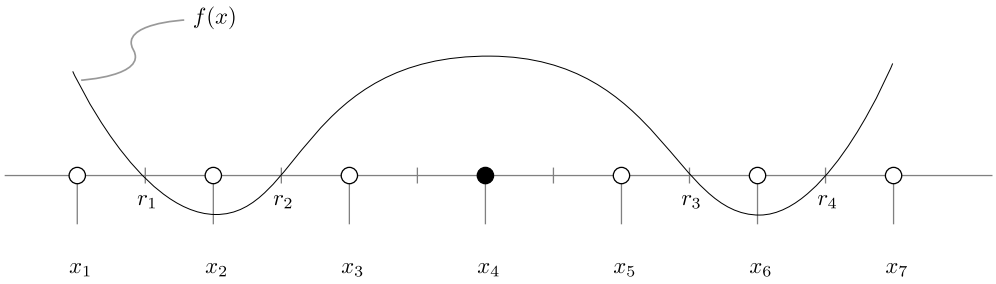


Figure C.2: Merge involving $n = 7$ points. f has in total $k = n - 3 = 4$ right roots.

Example 177 (Merge)**.** For merges (see Figure C.2), we only get right-roots and hence, we can write them (the merges) as sums of f_0 solutions. The polynomial has degree $n - 3$ (if the move involves n points) and so $\langle x \rangle = 0$, just as expected, for a merge.

This scheme fails for moves corresponding to lower bias Mochon's games. For instance, the bias $1/14$ move has its coordinates given by $x_0, x_1, x_2, x_3, x_4, x_5, x_6$ and the roots of f by l_1, l_2, r_1, r_2, r_3 which satisfy $x_0 < l_1 < l_2 < x_1 < x_2 \cdots < x_6 < r_1 < r_2 < r_3$. Here we can either consider l_1 to be a right-root, in which case l_2 is a left-root—a root which is not a right-root. Or we can consider l_2 to be a right-root, in which case l_1 becomes a left-root. Thus for Mochon's games with bias $1/14$ and less, we must revert to Lemma 82, which means we can not (at least by this scheme) avoid finding the solution to all the monomial assignments.

As we mentioned *merge*, for completeness we note that *split* is another counter-example. The situation (see Figure C.3) is similar to that of merge but with one key distinction: the polynomial has degree $n-2$; it has $n-3$ right-roots but 1 left-root (a root which is not a right-root). Thus, it too can not be expressed as a sum of f_0-assignments using Lemma 175. Of course, merges and splits by themselves are not of much interest in this discussion because we already know that the Blinkered Unitary solves both (see Subsection 4.2.1 of Chapter 6).

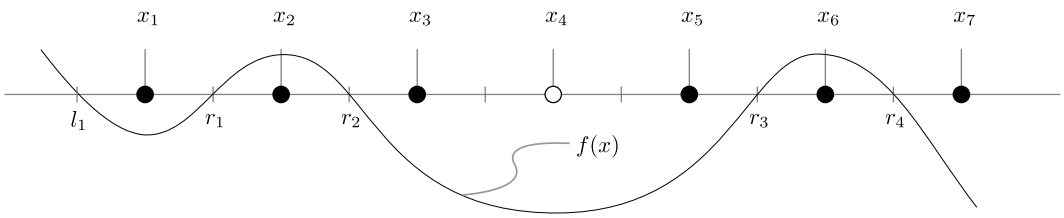


Figure C.3: Split involving 7 points. f has in total $k = n - 2 = 5$ roots (4 right and 1 left).

D

Appendix

Approaching $1/(4k+2)$ | Geometric Approach

In Chapter 5 we used some results which we now state and prove.

§ D.1 Known Results

Consider a curve in the plane specified by a function f. Its curvature is related to the rate of change of the tangents of f, i.e. the second derivative of f. For a surface in arbitrary dimensions specified by f, the corresponding quantity becomes a matrix $\partial_i \partial_j f$. The eigenvalues of this matrix tell us the curvature along the corresponding eigenvector. While in principle, it is possible to find this matrix by following this approach, in practice it becomes rather cumbersome[1]. Using a more general method one can easily obtain an analytic solution to this problem, for ellipsoids. The *Weingarten map*, defined intuitively, is the differential of the normal at a given point on the manifold. This turns out to be effectively the same as finding the aforementioned matrix of second derivatives.

Definition 178 (Weingarten Map (informal))**.** (see[2] § 2.5 of [33]) Let K be a manifold specified by the heads of vectors in $\mathbb{R}^n$. Denote the tangent space of K at $|x\rangle \in K$ by $T_{|x\rangle}K$. Let $|u_K\left(|x\rangle\right)\rangle$ be the outer unit normal vector of K at $|x\rangle$. The map $|u_k\left(|x\rangle\right)\rangle : K \to \mathbb{S}^{n-1} \subset \mathbb{R}^n$ as defined is called the *spherical image map* (or *Gauss map*) of the interior of the manifold K. Its differential at $|x\rangle$, $d(|u_K\rangle)_k =: W_x$ maps $T_{|x\rangle}K$ to itself. The linear map $W_x : T_{|x\rangle}K \to T_{|x\rangle}K$ is called the *Weingarten map*.

A related quantity, known as the *Reverse Weingarten map*, is easier to calculate. This is of interest because of the following result.

Theorem 179 (Informal)**.** *[33] The inverse of the Weingarten map equals the reverse Weingarten map, for well behaved surfaces.*

We omit the exact statement of the theorem and the definition of the Reverse Weingarten map as they are not directly relevant to the discussion. We simply work with a formula for the Weingarten map as described below.

Definition 180 (Support Function)**.** [33] Given a manifold specified by a set S of vectors, and a normalised vector $|u\rangle$, the support function is defined as

$$h_S(|u\rangle) := \sup_{|s\rangle \in S} \langle s|u\rangle \,.$$

[1] as one must choose a coordinate system with its origin at the point of interest, aligned along the normal and re-express all the quantities

[2] Note, their convention for T and K is slightly different; Informal because there are qualifying conditions on K which we suppressed.

Theorem 181 (Formula for evaluating the Reverse Weingarten Map (Informal)). *(see[3] § 2.5 of [33]) Consider a convex surface specified by a set S of vectors. Given a normalised vector $|u\rangle$, the reverse Weingarten map, W, evaluated along the normal specified by $|u\rangle$ is given by*

$$(W)_{ij} = \left.\frac{\partial^2 h_S(|u'\rangle)}{\partial u'_i \partial u'_j}\right|_u$$

where $h_S(|u'\rangle)$ is the support function.

Assuming that we can invert a matrix, using Theorem 181 and Theorem 179 one can obtain the Weingarten map. We apply this to the case of ellipsoids.

§ D.2 Normals and the Weingarten Map (Curvature)

Lemma 182. *(Also in Section 6.2) Given an $n \times n$ matrix $G \geq 0$, the support function corresponding to the ellipsoid S_G along a normal $|u\rangle$ of the manifold is given by*

$$h_{S_G}(|u\rangle) = \sqrt{\langle u| G^{\dashv} |u\rangle}.$$

Remark 183. Given an $n \times n$ matrix $G \geq 0$, note that $S_G = \{\mathcal{E}_G(|v\rangle) \mid \langle v|v\rangle = 1, |v\rangle \in \Pi\mathbb{R}^n\}$ where Π is as defined in Definition 145.

In our analysis, we typically know the point at which we wish to evaluate the curvature. The calculation of the support function requires the normal at that point. To this end, we give a formula for evaluating the latter.

Lemma 184 (Normal). *(Also in Section 6.2) Given an $n \times n$ matrix $G \geq 0$, consider the manifold S_G associated with it. Let $|v\rangle \in \Pi\mathbb{R}^n$ be a vector such that $\mathcal{E}_G(|v\rangle)$ is well-defined ($\langle v| G |v\rangle \neq 0$) where Π is as defined in Definition 145. The normal at $\mathcal{E}_G(|v\rangle)$ (which we also refer to as the normal along $|v\rangle$) is given by $|u\rangle = G|v\rangle / \sqrt{\langle v| G^2 |v\rangle}$.*

Proof. Consider the case where $G = \text{diag}(x_{g_1}, x_{g_2} \ldots x_{g_n})$ and let $|v\rangle = (v_1, v_2 \ldots v_n)$. The surface S_G is determined by the constraint $\langle v| G |v\rangle = 1$ which is equivalent to $\sum_{i=1}^n x_{g_i} v_i^2 = 1$. Changing the constant 1 can be thought of as scaling the surface. Treating $\sum_{i=1}^n x_{g_i} v_i^2$ as a scalar function, its gradient will point along the outward normal: $|u\rangle \propto \sum_{j=1}^n \frac{\partial}{\partial v_j} \sum_{i=1}^n x_{g_i} v_i^2 |j\rangle \propto \sum_{j=1}^n x_{g_j} v_j |j\rangle \propto G|v\rangle$. □

With these ingredients we can evaluate the Reverse Weingarten Map.

Lemma 185 (Reverse Weingarten Map). *Given an $n \times n$ matrix $G \geq 0$, and a vector $|v\rangle \in \Pi\mathbb{R}^n$ where Π is as defined in Definition 145, the reverse Weingarten Map associated with the surface S_G, evaluated at the point $\mathcal{E}_G(|v\rangle)$ is given by*

$$W_G := \sqrt{\frac{\langle G^2\rangle}{\langle G\rangle}} \left(G^{\dashv} - \frac{|v\rangle\langle v|}{\langle G\rangle}\right),$$

where $\langle G^j\rangle := \langle v| G^j |v\rangle$.

[3] Informal because the qualifying conditions on the surface and certain technicalities are missing.

Proof. We prove this for the case where $G > 0$ (the case when $G \geq 0$ but $G \not> 0$, follows analogously by restricting to the non-zero eigenspace). Let the spectral decomposition of G be given by

$$G = \sum_{i=1}^{n} x_{g_i} |g_i\rangle \langle g_i|$$

and let $|v\rangle = \sum_{i=1}^{n} c_i |g_i\rangle$. Recall that the normal along $|v\rangle$ (see Lemma 184) is given by $|u\rangle = G|v\rangle / \sqrt{\langle v| G^2 |v\rangle}$. Writing $|u\rangle = \sum_{i=1}^{n} u_i |g_i\rangle$, u_is are fixed. Then the support function evaluated along the normal $|u\rangle$ is given by (we use h to denote $h_{S_G}(|u\rangle)$ for brevity)

$$h = \sqrt{\langle u| G^{\dashv} |u\rangle} = \sqrt{\sum_{i=1}^{n} x_{g_i}^{-1} u_i^2} \qquad \text{using Lemma 182}$$

$$\implies (W_G)_{ij} = \frac{\partial^2 h}{\partial u_i \partial u_j} = -\frac{1}{h^3} x_{g_j}^{-1} x_{g_i}^{-1} u_j u_i + \frac{x_{g_i}^{-1}}{h} \delta_{ij} \qquad \text{using Theorem 181}$$

$$\implies W_G = -\frac{1}{h^3} G^{\dashv} |u\rangle \langle u| G^{\dashv} + \frac{G^{\dashv}}{h}$$

where we used the more general notation $G^{\dashv} = G^{-1}$ (because in our case $G > 0$). Substituting $|u\rangle$ in the expression for h and for W_G we obtain the required result, viz. $h = \sqrt{\frac{\langle G \rangle}{\langle G^2 \rangle}}$, and

$$W_G = \frac{1}{h} G^{\dashv} - \frac{1}{h^3} \frac{|v\rangle \langle v|}{\langle G^2 \rangle} = \sqrt{\frac{\langle G^2 \rangle}{\langle G \rangle}} G^{\dashv} - \frac{\cancel{\langle G^2 \rangle} \sqrt{\langle G^2 \rangle}}{\langle G \rangle \sqrt{\langle G \rangle}} \frac{|v\rangle \langle v|}{\cancel{\langle G^2 \rangle}} = \sqrt{\frac{\langle G^2 \rangle}{\langle G \rangle}} \left(G^{\dashv} - \frac{|v\rangle \langle v|}{\langle G \rangle} \right).$$

When $G \geq 0$ and has zero eigenvalues, the spectral decomposition has m elements with $m < n$, viz. $G = \sum_{i=1}^{m} x_{g_i} |g_i\rangle \langle g_i|$. The sum $\sum_{i=1}^{m} x_{g_i}^{-1} u_i^2$ would then correspond to $\langle u| G^{\dashv} |u\rangle$. Similar replacements can be made to generalise the proof for $G \geq 0$. □

Inverting the Reverse Weingarten Map is also not too hard due to the following result. Combining these, we obtain the Weingarten map.

Theorem 186 (Sherman-Morrison formula). *[34, 19] Let A be an $n \times n$ invertible matrix and let $|a\rangle, |b\rangle$ be vectors (in n-dimensions). Then, $(A + |a\rangle \langle b|)$ is invertible if and only if $1 + \langle b| A^{-1} |a\rangle \neq 0$. Further, if this is the case, then*

$$(A + |a\rangle \langle b|)^{-1} = A^{-1} - \frac{A^{-1} |a\rangle \langle b| A^{-1}}{1 + \langle b| A^{-1} |a\rangle}.$$

Lemma 187 (Weingarten Map). *Given an $n \times n$ matrix $G \geq 0$, the Weingarten Map associated with the surface S_G, evaluated at the point $\mathcal{E}_G(|v\rangle)$ is given by*

$$W_G^{\dashv} = \sqrt{\frac{\langle G \rangle}{\langle G^2 \rangle}} \left(G + \frac{\langle G^3 \rangle}{\langle G^2 \rangle^2} G |v\rangle \langle v| G - \frac{1}{\langle G^2 \rangle} \left(G |v\rangle \langle v| G^2 + G^2 |v\rangle \langle v| G \right) \right)$$

where $\langle G \rangle := \langle v| G |v\rangle$.

Proof. Again, we prove this for the case $G > 0$ and the proof for the case where $G \geq 0$ follows analogously. By a direct computation, it is clear that $W_G G |v\rangle = 0$ (see Lemma 185), and applying Theorem 186 we obtain

$$W^{-1} = \sqrt{\frac{\langle G \rangle}{\langle G^2 \rangle}} \left(G + \frac{G |v\rangle \langle v| G}{\langle G \rangle \cdot 0} \right), \tag{D.1}$$

where we set $A = G^{\dashv} = G^{-1}$ (in this case) and $|a\rangle = |b\rangle = G|v\rangle / \sqrt{\langle G\rangle}$ (after pulling out the $1/\sqrt{\langle G^2\rangle / \langle G\rangle}$ factor). Using appropriate interpolations (for instance one could use $|a\rangle = -|b\rangle = (1-\epsilon)G|v\rangle / \sqrt{\langle G\rangle}$ instead of $G|v\rangle / \sqrt{\langle G\rangle}$), one can make the second term well behaved and have it diverge only as some parameter vanishes ($\epsilon = 0$). The quantity we are interested in is $W^{\dashv} = \Pi_u^{\perp} W \Pi_u^{\perp}$, where $\Pi_u^{\perp} = \mathbb{I} - |u\rangle\langle u|$ and $|u\rangle = G|v\rangle / \sqrt{\langle G\rangle}$. If the positive inverse is to be well-defined, the second term in Equation (D.1) should disappear after the projection, viz. $\Pi_u^{\perp} G|v\rangle\langle v|G\Pi_u^{\perp}$ should vanish. Indeed, it does because $G|v\rangle \propto |u\rangle$. The non-vanishing contribution must then come from the first term in Equation (D.1), $\Pi_u^{\perp} G \Pi_u^{\perp} = (\mathbb{I} - |u\rangle\langle u|)\, G\, (\mathbb{I} - |u\rangle\langle u|)$, which entails

$$
\begin{aligned}
W^{\dashv} &= \sqrt{\frac{\langle G\rangle}{\langle G^2\rangle}} \Pi_u^{\perp} G \Pi_u^{\perp} = \sqrt{\frac{\langle G\rangle}{\langle G^2\rangle}} \left(G - G|u\rangle\langle u| - |u\rangle\langle u|G + \langle u|G|u\rangle |u\rangle\langle u|\right) \\
&= \sqrt{\frac{\langle G\rangle}{\langle G^2\rangle}} \left(G - \frac{G^2|v\rangle\langle v|G}{\langle G^2\rangle} - \frac{G|v\rangle\langle v|G^2}{\langle G^2\rangle} + \frac{\langle G^3\rangle}{\langle G^2\rangle}\frac{G|v\rangle\langle v|G}{\langle G^2\rangle}\right).
\end{aligned}
$$

The case for $G \geq 0$ where G has zero eigenvalues carries through. This can be seen by viewing the Sherman Morrison formula as a "correction" to an inverse when one entry of the matrix is changed. The inverse of G we are interested in is the positive inverse $G^{\dashv}$. The entry of the matrix that we change is in this positive subspace. Restricting the analysis to this subspace, the matrix G can be viewed as positive, viz. $G > 0$, yielding the required generalisation. □

§ D.3 The $-1/x$ trick and monomial assignments

In Subsection 7.6.1, we considered a monomial assignment with the highest permissible degree (of the monomial) and asserted that this was effectively an f_0 assignment. We used this again in Subsection 7.6.2. Here, we prove this assertion. We use the following result–a relation between an EBRM function (an EBM function with the matrices restricted to be real) with the spectra of their matrices in $[\chi, \xi]$ and valid functions with support in $[\chi, \xi]$. We proved it in Section 6.1 of Chapter 6.

Lemma 188. *(Restatement of Lemma 130) A function $t = \sum_i p_i [\![x_i]\!]$ is EBRM on $[\chi, \xi]$ if and only if it is $[\chi, \xi]$-valid (corresponds to requiring $\sum_i p_i f_\lambda(x_i) \geq 0$ for all $\lambda \in (-\infty, \infty)\backslash[-\xi, -\chi]$ with $f_\lambda(x) = -1/(\lambda + x)$).*

This is of interest to us because it lets us replace $[\![x_i]\!]$ with $[\![1/x_i]\!]$ at the cost of a minus sign. Mochon's f-assignments have a structure which transforms in a useful way under $x_i \mapsto 1/x_i$. We can combine these to show that monomial and effectively monomial assignments (see Corollary 190) are equally easy to solve; if O solves one, O^T solves the other. Further, we can use the transformation, $[\![x_i]\!] \mapsto -[\![1/x_i]\!]$, to convert left-roots into right-roots. Later, we combine these to show that any f-assignment can be expressed as a sum of monomial assignments and/or effectively monomial assignments. We now state and prove these statements.

Lemma 189. *Let $\chi, \xi > 0$. A function $t = \sum_i p_i [\![x_i]\!]$ is $[\chi, \xi]$-EBRM if and only if $t' = \sum_i -p_i [\![1/x_i]\!]$ is $[1/\xi, 1/\chi]$-EBRM. Further, if O solves the matrix instance corresponding to t with spectrum in $[\chi, \xi]$ then O^T solves that of t' with spectrum in $[1/\xi, 1/\chi]$.*

Proof. We start with the only if part ($\implies$). We are given H, G with spectrum in $[\chi, \xi]$ and a vector $|w\rangle$ such that $t = \text{Prob}[H, |w\rangle] - \text{Prob}[G, |w\rangle]$ and $H \geq G$. Further, $H \geq G \iff H^{-1} \leq G^{-1}$. By using spectral decomposition, one should be able to see that $t' = \text{Prob}[G^{-1}, |w\rangle] - \text{Prob}[H^{-1}, |w\rangle]$.

Defining $H' = G^{-1}, G' = H^{-1}, |w'\rangle = |w\rangle$, we have $t' = \text{Prob}[H', |w'\rangle] - \text{Prob}[G', |w'\rangle]$ and $H' \geq G'$ where G' and H' have their spectrum in $[1/\xi, 1/\chi]$. The same argument should also work for the other direction ($\Longleftarrow$). The last statement is seen to be true by using a basis in which $H = X_h$ is diagonal, writing $G = OX_gO^T$ and noting that $O^{-1} = O^T$. □

Corollary 190 (Effectively monomial assignment). *Let* $0 < x_1 < x_2 \cdots < x_n$. *Then,* O^T *solves a matrix instance corresponding to*

$$t = \sum_{i=1}^{n} \frac{\left(-\frac{1}{x_i}\right)^k}{\prod_{i\neq j}(\frac{1}{x_j} - \frac{1}{x_i})} [\![x_i]\!]$$

if and only if O *solves the corresponding matrix instance associated with the monomial assignment*

$$t' = \sum_{i=1}^{n} \frac{-\left(-\frac{1}{x_i}\right)^k}{\prod_{i\neq j}(\frac{1}{x_j} - \frac{1}{x_i})} \left[\!\!\left[\frac{1}{x_i}\right]\!\!\right] = \sum_{i=1}^{n} \frac{-(-\omega_i)^k}{\prod_{i\neq j}(\omega_j - \omega_i)} [\![\omega_i]\!]$$

where $\omega_i := 1/x_i$. *We therefore refer to* t *as an* effectively monomial assignment.

§ D.4 Existence of Solutions to Matrix Instances and their dimensions

The goal of this discussion is to show that certain matrix instances (corresponding to Mochon's monomial assignments) can be solved with low dimensional matrices. The argument is essentially the same as the one used in the EMA algorithm (see the discussion before Subsection 6.3.3.3, in Subsection 1.3.3, of Chapter 6). We explain it again in the notation we used in Chapter 5.

From Lemma 191 and Lemma 192 (see below) we know that a solution to a matrix instance corresponding to a $[\chi, \xi]$ valid function always exists, granted we pad the matrices with χ and ξ to have their size equal to $n \times n$ with $n = n_g + n_h - 1$. We can however do even better. To see this, consider the matrix instance $\underline{\mathrm{X}}^{\bar{k}}$ in the notation introduced in Lemma 192. The eigenspace of H on which $|w\rangle$ has a component, is of size n_h (similarly with G, $|v\rangle$ and n_g). Every time we iterate using the Weingarten map, we remove one component from both H and $|w\rangle$ from within this eigenspace (similarly for G and $|v\rangle$). Consequently, in the subsequent step, the eigenspace of $H^{\overline{k-1}}$ on which $\left|w^{\overline{k-1}}\right\rangle$ has a component, is of size $n_h - 1$ (similarly with $G^{\overline{k-1}}$, $\left|v^{\overline{k-1}}\right\rangle$ the size becomes $n_g - 1$) where the matrix instance after the Weingarten Iteration map was taken to be $\underline{\mathrm{X}}^{\overline{k-1}} =: (H^{\overline{k-1}}, G^{\overline{k-1}}, \left|w^{\overline{k-1}}\right\rangle, \left|v^{\overline{k-1}}\right\rangle)$. In case of the balanced f_0 assignment, we end up with a matrix instance $\underline{\mathrm{X}}^{l} =: (H^{\dot{l}}, G^{\dot{l}}, 0, 0)$ where the vectors disappear. The matrices $H^{\dot{l}}$ and $G^{\dot{l}}$ only have ξ and χ, respectively, as their eigenvalues and then, we trivially have $H^{\dot{l}} > G^{\dot{l}}$. In fact, this part of the matrix plays no role and can be removed. This justifies why we could assume that even without padding with χs and ξs, the matrix instance corresponding to the f_0 assignment had a solution.

The padding becomes important, however, when we use the Wiggle-w (or Wiggle-v) map to iterate. To see this, consider again the matrix instance $\underline{\mathrm{X}}^{\bar{k}}$ in the notation introduced in Lemma 192 (ξ would tend to ∞ in these cases). The eigenspace of H on which $|w\rangle$ has a component, is of size n_h. Every time we iterate using the Wiggle-w map, we effectively do not remove any component from H and $|w\rangle$ from within this eigenspace. This is because we introduce an extra dimension (in our discussions, we formalised it as H having wiggle-w room along $|t_h\rangle$; here it can be thought of as any one of the

$|h_i\rangle$s with $i > n_h$), and then we project out one dimension, leaving the overall dimension of the space unchanged. The dimension for the G and $|v\rangle$ case, however, drops as before. Again, when we reach a matrix instance $\underline{X}^j$ (after applying a combination of Weingarten Iteration maps, Wiggle-v/w Iteration maps), where the vectors disappear we can use the reasoning above to justify that matrices with fewer padded dimensions also have a solution.

Lemma 191. *(Restatement of Lemma 104) Let $t = h - g = \sum_{i=1}^{m} p_i [\![x_i]\!]$ be a $[\chi, \xi]$ valid function where $h =: \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!]$ and $g =: \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$ have disjoint support and $p_{h_i} > 0$ and $p_{g_i} > 0$ (for $i \in \{1, 2 \dots n_h\}$ and $\{1, 2 \dots n_g\}$ respectively). Let X_h and X_g be $n \times n$ diagonal matrices, where $n = n_h + n_g - 1$, given by*

$$X_h = diag(x_{h_1}, x_{h_2}, \dots x_{h_{n_h}}, \xi, \xi \dots \xi)$$
$$X_g = diag(x_{g_1}, x_{g_2}, \dots x_{g_{n_g}}, \chi, \chi \dots \chi).$$

Then there exists an orthogonal matrix O which solves the matrix instance $\underline{X}^{\bar{n}} := (X_h, X_g, |w\rangle, |v\rangle)$.

Lemma 192. *Let k, n_h and n_g be strictly positive integers such that $k \geq n_h$ and $k \geq n_g$. Consider a matrix instance $\underline{X}^{\bar{k}} =: (H, G, |w\rangle, |v\rangle)$ where*

$$H = \sum_{i=1}^{n_h} x_{h_i} |h_i\rangle \langle h_i| + \sum_{i=n_h+1}^{k} \xi |h_i\rangle \langle h_i|$$
$$|w\rangle = \sum_{i=1}^{n_h} \sqrt{p_{h_i}} |h_i\rangle$$

and

$$G = \sum_{i=1}^{n_g} x_{g_i} |g_i\rangle \langle g_i| + \sum_{i=n_g+1}^{k} \chi |g_i\rangle \langle g_i|$$
$$|v\rangle = \sum_{i=1}^{n_g} \sqrt{p_{g_i}} |g_i\rangle$$

such that

$$x_{h_i} \neq x_{g_j},$$
$$p_{h_i} > 0,$$
$$p_{g_j} > 0$$

hold for all $i \in \{1, 2 \dots n_h\}$, $j \in \{1, 2 \dots n_g\}$, and $\mathcal{H}^{\bar{k}} = span\{|h_i\rangle\}$, $\mathcal{G}^{\bar{k}} = span\{|g_i\rangle\}$ (see Definition 155). Then if the isometry $Q : \mathcal{H}^{\bar{k}} \to \mathcal{G}^{\bar{k}}$ solves the matrix instance $\underline{X}^{\bar{k}}$ then the function

$$t = \sum_{i=1}^{n_h} p_{h_i} [\![x_{h_i}]\!] - \sum_{i=1}^{n_g} p_{g_i} [\![x_{g_i}]\!]$$

is $[\chi, \xi]$-valid (which is equivalent to being $[\chi, \xi]$-EBRM).

§ D.5 Lemmas for the Contact and Component conditions

Lemma 193. *Consider the matrix instance* $\underline{X}^{\bar{n}} := (H^{\bar{n}}, G^{\bar{n}}, \left|w^{\bar{n}}\right\rangle, \left|v^{\bar{n}}\right\rangle)$. *Suppose that the Weingarten Iteration Map (see Definition 157) is applied* l *times to obtain* $\underline{X}^{\overline{n-l}} := \left(H^{\overline{n-l}}, G^{\overline{n-l}}, \left|w^{\overline{n-l}}\right\rangle, \left|v^{\overline{n-l}}\right\rangle\right)$. *Then, for any* l, *the expectation value* $\left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle$ *is a function of the expectation values* $\left\langle v^{\bar{n}}\right| (G^{\bar{n}})^p \left|w^{\bar{n}}\right\rangle = \langle (G^{\bar{n}})^p \rangle$, *where the powers* p *range from* 0 *to* $2l+1$ *at most. The corresponding statement involving* H*'s and* $|w\rangle$*'s also holds.*

Proof. Using once the Weingarten Iteration Map, we obtain:

$$\begin{aligned}
\left|v^{\overline{n-1}}\right\rangle &= \left|v^{\bar{n}}\right\rangle - \frac{\langle G^{\bar{n}}\rangle}{\langle (G^{\bar{n}})^2\rangle} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \\
G^{\overline{n-1}} &= G^{\bar{n}} + \frac{\langle (G^{\bar{n}})^3\rangle}{\langle (G^{\bar{n}})^2\rangle^2} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} \\
&\quad - \frac{1}{\langle (G^{\bar{n}})^2\rangle} \left(G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^2 + (G^{\bar{n}})^2 \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} \right).
\end{aligned} \tag{D.2}$$

If we continue to iterate accordingly and express everything in terms of $\left|v^{\bar{n}}\right\rangle$ and $G^{\bar{n}}$, which are known, after l steps we will obtain:

$$\begin{aligned}
\left|v^{\overline{n-l}}\right\rangle &= \sum_{i=0}^{l} \alpha_i (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \\
G^{\overline{n-l}} &= G^{\bar{n}} + \sum_{i,j=0}^{l+1} \alpha_{i,j} (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j,
\end{aligned} \tag{D.3}$$

where the multiplicative factors a_i and $a_{i,j}$ also contain terms of the form $\langle (G^{\bar{n}})^p \rangle$, in which p ranges between the minimum and maximum powers appearing in the sum (see remark at the end of the proof).

Indeed, we can use induction to prove that Equation (D.3) holds for all l.
The base of the induction $l = 1$ immediately gives us Equation (D.2).
For the $l+1$ instance, using the Weingarten Iteration Map, we have:

$$\begin{aligned}
\left|v^{\overline{n-l-1}}\right\rangle &= \left|v^{\overline{n-l}}\right\rangle - \frac{\langle G^{\overline{n-l}}\rangle}{\langle (G^{\overline{n-l}})^2\rangle} (G^{\overline{n-l}}) \left|v^{\overline{n-l}}\right\rangle \\
G^{\overline{n-l-1}} &= G^{\bar{n}} + \frac{\langle (G^{\overline{n-l}})^3\rangle}{\langle (G^{\overline{n-l}})^2\rangle^2} G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle \left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \\
&\quad - \frac{1}{\langle (G^{\overline{n-l}})^2\rangle} \left(G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle \left\langle v^{\overline{n-l}}\right| (G^{\overline{n-l}})^2 + (G^{\overline{n-l}})^2 \left|v^{\overline{n-l}}\right\rangle \left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \right).
\end{aligned}$$

Replacing $G^{\overline{n-l}}$ and $\left|v^{\overline{n-l}}\right\rangle$ from Equation (D.3), we get

$$\begin{aligned}
\left|v^{\overline{n-l-1}}\right\rangle &= \textstyle\sum_{i=0}^{l+1} \alpha_i (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \\
G^{\overline{n-l-1}} &= G^{\bar{n}} + \textstyle\sum_{i,j=0}^{l+2} \alpha_{i,j} (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j,
\end{aligned} \tag{D.4}$$

which proves that Equation (D.3) is valid for all l.
We can now complete our proof by expressing $\left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle$ in terms of $\langle G^{\bar{n}} \rangle$. Substituting from

Equation (D.3), we get:

$$\left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle = \sum_{i=0}^{l} \alpha_i \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{i+1} \sum_{j=0}^{l} \alpha_j (G^{\bar{n}})^{j} \left|v^{\bar{n}}\right\rangle \tag{D.5}$$
$$+ \sum_{i=0}^{l} \alpha_i \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{i} \sum_{i',j'=0}^{l+1} \alpha_{i',j'} (G^{\bar{n}})^{i'} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{j'} \sum_{j=0}^{l} \alpha_j (G^{\bar{n}})^{j} \left|v^{\bar{n}}\right\rangle .$$

In Equation (D.5), we see that the minimum expectation value is $\langle (G^{\bar{n}})^0 \rangle$, while the maximum is $\langle (G^{\bar{n}})^{2l+1} \rangle$, which concludes the proof. □

Notice that we left a_i and $a_{i,j}$ undetermined and we even used the same notation for them (obviously a_i and $a_{i,j}$ are different in Equation (D.3), Equation (D.4) and Equation (D.5)). In the context of our proof their specific form is not relevant, but what is rather important are the minimum and maximum powers p in $\langle (G^{\bar{n}})^p \rangle$ that contain and might appear in $\left\langle v^{\overline{n-l}}\right| G^{\overline{n-l}} \left|v^{\overline{n-l}}\right\rangle$. To estimate them, it suffices to observe that the minimum power in $\left|v^{\overline{n-l}}\right\rangle$ comes from the first term $\left|v^{\bar{n}}\right\rangle$ and is 0, while the maximum power that appears in $\left|v^{\overline{n-l}}\right\rangle$ comes from $\langle (G^{\overline{n-l+1}})^2 \rangle$ (see Definition 157) and is equal to $2l$. In $G^{\overline{n-l}}$, however, we can find an even higher power appearing in the $a_{i,j}$'s coming from $\langle (G^{\overline{n-l+1}})^3 \rangle$ (see Definition 157) and is equal to $2l+1$. In total these powers are always between the minimum and maximum powers on Equation (D.5), thus the factors a_i and $a_{i,j}$ do not need to be specified.

Lemma 194. *Consider the extended matrix instance* $\underline{M}^{\bar{n}} := \mathscr{U}(H^{\bar{n}}, G^{\bar{n}}, \left|w^{\bar{n}}\right\rangle, \left|v^{\bar{n}}\right\rangle, (H^{\bar{n}})^{\dashv}, (G^{\bar{n}})^{\dashv}, |.\rangle, |.\rangle)$. *Suppose the Normal Initialisation Map and the Weingarten Iteration Map (see Definition 156 and Definition 157) are applied* l *times to obtain* $\underline{M}^{\overline{n-l}}$*, viz. applying* $\underline{M}^{\overline{i-1}} = \mathscr{U}(\mathscr{W}(\underline{M}^{i}))$ l *times. Then, for any* l*, the expectation value* $\left\langle v^{\overline{n-l}}\right| (G^{\overline{n-l}})^{\dashv} \left|v^{\overline{n-l}}\right\rangle$ *is a function of the expectation values* $\left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{p} \left|w^{\bar{n}}\right\rangle = \langle (G^{\bar{n}})^{p} \rangle$*, where the powers* p *range from* 0 *to* $2l+1$ *at most. The corresponding statement involving* H*'s and* $|w\rangle$*'s also holds.*

Proof. First, we need to specify the form of $(G^{\overline{n-l}})^{\dashv}$ as a function of $G^{\bar{n}}$ and $\left|v^{\bar{n}}\right\rangle$. The first iteration gives:

$$\begin{aligned} \left|v^{\overline{n-1}}\right\rangle &= \left|v^{\bar{n}}\right\rangle - \frac{< G^{\bar{n}} >}{\langle (G^{\bar{n}})^2 \rangle} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \\ (G^{\overline{n-1}})^{\dashv} &= (G^{\bar{n}})^{\dashv} - \frac{\left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right|}{\langle G^{\bar{n}} \rangle}. \end{aligned} \tag{D.6}$$

Continuing the iterations to l, we obtain:

$$\begin{aligned} \left|v^{\overline{n-l}}\right\rangle &= \sum_{i=0}^{l} \alpha_i (G^{\bar{n}})^{i} \left|v^{\bar{n}}\right\rangle \text{ (from the previous lemma)} \\ (G^{\overline{n-l}})^{\dashv} &= (G^{\bar{n}})^{\dashv} + \sum_{i,j=0}^{l-1} \alpha_{i,j} (G^{\bar{n}})^{i} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{j}. \end{aligned} \tag{D.7}$$

Indeed, by induction we can prove that Equation (D.7) holds for all l.
The base of the induction $l = 1$ immediately gives us Equation (D.6), which holds.

For the $l+1$ instance, the Weingarten Iteration Map gives us:

$$\begin{aligned}\left|v^{\overline{n-l-1}}\right\rangle &= \left|v^{\overline{n-l}}\right\rangle - \frac{\langle G^{\overline{n-l}}\rangle}{\langle (G^{\overline{n-l}})^2\rangle}(G^{\overline{n-l}})\left|v^{\overline{n-l}}\right\rangle \\ (G^{\overline{n-l-1}})^{\dashv} &= (G^{\overline{n-l}})^{\dashv} - \frac{\left|v^{\overline{n-l}}\right\rangle\left\langle v^{\overline{n-l}}\right|}{\langle G^{\overline{n-l}}\rangle}.\end{aligned} \tag{D.8}$$

Replacing $G^{\overline{n-l}}$ and $\left|v^{\overline{n-l}}\right\rangle$ from Equation (D.7), we get

$$\begin{aligned}\left|v^{\overline{n-l-1}}\right\rangle &= \textstyle\sum_{i=0}^{l+1}\alpha_i (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \\ (G^{\overline{n-l-1}})^{\dashv} &= (G^{\bar{n}}) + \textstyle\sum_{i,j=0}^{l}\alpha_{i,j}(G^{\bar{n}})\left|v^{\bar{n}}\right\rangle\left\langle v^{\bar{n}}\right|(G^{\bar{n}})^j,\end{aligned} \tag{D.9}$$

which concludes our inductive proof.

Now that we proved that Equation (D.7) holds for any l, we can proceed to the calculation of the corresponding expectation value:

$$\begin{aligned}\langle (G^{\overline{n-l}})^{\dashv}\rangle \;=\; & \left\langle v^{\overline{n-l}}\right|(G^{\overline{n-l}})\left|v^{\overline{n-l}}\right\rangle = \sum_{i,j=0}^{l}\alpha_i\alpha_j\left\langle v^{\bar{n}}\right|(G^{\bar{n}})^{i+j-1}\left|v^{\bar{n}}\right\rangle \\ & + \sum_{i=0}^{l}\left\langle v^{\bar{n}}\right|(G^{\bar{n}})^i\sum_{i',j'=0}^{l-1}\alpha_{i',j'}(G^{\bar{n}})^{i'}\left|v^{\bar{n}}\right\rangle\left\langle v^{\bar{n}}\right|(G^{\bar{n}})^{j'}\sum_{j=0}^{l}\alpha_j(G^{\bar{n}})^j\left|v^{\bar{n}}\right\rangle,\end{aligned} \tag{D.10}$$

where we have used $(G^{\bar{n}})^{\dashv} = (G^{\bar{n}})^{-1}$, since $G^{\bar{n}}$ is full rank.
We observe that the minimum power in the expectation value is $\langle G^{\bar{n}}\rangle$, while the maximum is $\langle (G^{\bar{n}})^{2l-1}\rangle$. Recall though (from the previous lemma) that in the multiplicative factors α_i and $\alpha_{i,j}$ there are higher powers in the expectation values $\langle (G^{\bar{n}})^{2l+1}\rangle$, which from now on will be the highest. Since we are iterating with respect to $G^{\dashv}$ the powers are not growing any more, but they rather decrease and we are interested on the minimum powers that are reduced with each iteration. □

Lemma 195. *Consider the extended matrix instance* $\underline{\tilde{M}}^{\bar{n}} := \mathscr{U}((H^{\bar{n}})^{\dashv}, (G^{\bar{n}})^{\dashv}, \left|\tilde{w}^{\bar{n}}\right\rangle, \left|\tilde{v}^{\bar{n}}\right\rangle, H^{\bar{n}}, G^{\bar{n}}, |.\rangle, |.\rangle)$. *Suppose the Normal Initialisation Map and the Weingarten Iteration Map (see Definition 156 and Definition 157) are applied* k *times to obtain* $\underline{\tilde{M}}^{\overline{n-k}}$, *viz. applying* $\underline{\tilde{M}}^{\overline{i-1}} = \mathscr{U}(\mathscr{W}(\underline{\tilde{M}}^{i}))$ k *times. Then, for any* k, *the expectation value* $\left\langle \tilde{v}^{\overline{d-k}}\right|\tilde{G}^{\overline{d-k}}\left|\tilde{v}^{\overline{d-k}}\right\rangle$ *is a function of the expectation values* $\left\langle v^{\bar{n}}\right|(G^{\bar{n}})^p\left|w^{\bar{n}}\right\rangle = \langle (G^{\bar{n}})^p\rangle$, *where the minimum power* p *that might appear is* $-(2k+1)$. *The corresponding statement involving* H*'s and* $|w\rangle$*'s also holds.*

Proof. The first iteration gives:

$$\begin{aligned}\left|\tilde{v}^{\overline{d-1}}\right\rangle \;=\; & \left|\tilde{v}^{\bar{d}}\right\rangle - \frac{\langle \tilde{G}^{\bar{d}}\rangle}{\langle (\tilde{G}^{\bar{d}})^2\rangle}\tilde{G}^{\bar{d}}\left|\tilde{v}^{d}\right\rangle \\ \tilde{G}^{\overline{d-1}} \;=\; & \tilde{G}^{\bar{d}} + \frac{\langle (\tilde{G}^{\bar{d}})^3\rangle}{\langle (\tilde{G}^{\bar{d}})^2\rangle^2}\tilde{G}^{\bar{d}}\left|\tilde{v}^{\bar{d}}\right\rangle\left\langle \tilde{v}^{\bar{d}}\right|\tilde{G}^{\bar{d}} \\ & - \frac{1}{\langle (\tilde{G}^{\bar{d}})^2\rangle}\left(\tilde{G}^{\bar{d}}\left|\tilde{v}^{\bar{d}}\right\rangle\left\langle \tilde{v}^{\bar{d}}\right|(\tilde{G}^{\bar{d}})^2 + (\tilde{G}^{\bar{d}})^2\left|\tilde{v}^{\bar{d}}\right\rangle\left\langle \tilde{v}^{\bar{d}}\right|\tilde{G}^{\bar{d}}\right).\end{aligned} \tag{D.11}$$

Continuing for k iterations, we can prove by induction that:

$$\begin{aligned}\left|\tilde{v}^{\overline{d-k}}\right\rangle &= \sum_{i=0}^{k} \alpha_i (G^{\bar{n}})^{i-k} \left|v^{\bar{n}}\right\rangle \\ \tilde{G}^{\overline{d-k}} &= (G^{\bar{n}})^{\dashv} + \sum_{i,j=0}^{k} \alpha_{i,j} (G^{\bar{n}})^{i-(k+1)} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{j-(k+1)}.\end{aligned} \tag{D.12}$$

Indeed, the base of the induction $k = 1$ gives us Equation (D.11), which holds.

For $k+1$, we obtain:

$$\begin{aligned}\left|\tilde{v}^{\overline{d-k-1}}\right\rangle &= \left|\tilde{v}^{\overline{d-k}}\right\rangle - \frac{\langle \tilde{G}^{\overline{d-k}} \rangle}{\langle (\tilde{G}^{\overline{d-k}})^2 \rangle} \tilde{G}^{\overline{d-k}} \left|\tilde{v}^{\overline{d-k}}\right\rangle \\ \tilde{G}^{\overline{d-k-1}} &= \tilde{G}^{\overline{d-k}} + \frac{\langle (\tilde{G}^{\overline{d-k}})^3 \rangle}{\langle (\tilde{G}^{\overline{d-k}})^2 \rangle^2} \tilde{G}^{\overline{d-k}} \left|\tilde{v}^{\overline{d-k}}\right\rangle \left\langle \tilde{v}^{\overline{d-k}}\right| \tilde{G}^{\overline{d-k}} \\ &\quad - \frac{1}{\langle (\tilde{G}^{\overline{d-k}})^2 \rangle} \left(\tilde{G}^{\overline{d-k}} \left|\tilde{v}^{\overline{d-k}}\right\rangle \left\langle \tilde{v}^{\overline{d-k}}\right| (\tilde{G}^{\overline{d-k}})^2 + (\tilde{G}^{\overline{d-k}})^2 \left|\tilde{v}^{\overline{d-k}}\right\rangle \left\langle \tilde{v}^{\overline{d-k}}\right| \tilde{G}^{\overline{d-k}} \right).\end{aligned}$$

Substituting $\left|\tilde{v}^{\overline{d-k}}\right\rangle$ and $\tilde{G}^{\overline{d-k}}$ from Equation (D.12), we get:

$$\begin{aligned}\left|\tilde{v}^{\overline{d-k-1}}\right\rangle &= \sum_{i=0}^{k+1} \alpha_i (G^{\bar{n}})^{i-k-1} \left|v^{\bar{n}}\right\rangle \\ \tilde{G}^{\overline{d-k-1}} &= (G^{\bar{n}})^{\dashv} + \sum_{i,j=0}^{k+1} \alpha_{i,j} (G^{\bar{n}})^{i-k-2} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{j-k-2},\end{aligned} \tag{D.13}$$

which confirms that Equation (D.12) holds for all k.

Thus, for any k the corresponding expectation value can be written as:

$$\begin{aligned}&\left\langle \tilde{v}^{\overline{d-k}}\right| \tilde{G}^{\overline{d-k}} \left|\tilde{v}^{\overline{d-k}}\right\rangle = \\ &\sum_{i=0}^{k} \alpha_i \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{i-k} (G^{\bar{n}})^{-1} \sum_{j=0}^{k} \alpha_j (G^{\bar{n}})^{j-k} \left|v^{\bar{n}}\right\rangle \\ &+ \sum_{i=0}^{l+k} \alpha_i \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{i-k} \sum_{i',j'=0}^{k} \alpha_{i',j'} (G^{\bar{n}})^{i'-(k+1)} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^{j'-(k+1)} \sum_{j=0}^{k} \alpha_j (G^{\bar{n}})^{j-k} \left|v^{\bar{n}}\right\rangle.\end{aligned}$$

We observe that the minimum power that can appear in the expectation values is $-(2k+1), \forall k$. Recall that the multiplicative factors α_i and $\alpha_{i,j}$, also contain terms of the form $\langle (G^{\bar{n}})^p \rangle$, which behave as explained in the previous lemmas. □

Lemma 196. *Consider the matrix instance* $\underline{X}^{\bar{n}} := (H^{\bar{n}}, G^{\bar{n}}, \left|w^{\bar{n}}\right\rangle, \left|v^{\bar{n}}\right\rangle)$. *Using the Weingarten Iteration Map once, we obtain:*

$$\begin{aligned}
\left|v^{\overline{n-1}}\right\rangle = &\quad \left|v^{\bar{n}}\right\rangle - \frac{\langle G^{\bar{n}}\rangle}{\langle (G^{\bar{n}})^2\rangle} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \\
G^{\overline{n-1}} = &\quad G^{\bar{n}} + \frac{\langle (G^{\bar{n}})^3\rangle}{\langle (G^{\bar{n}})^2\rangle^2} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} - \\
&\frac{1}{\langle (G^{\bar{n}})^2\rangle} \left(G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^2 + (G^{\bar{n}})^2 \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} \right).
\end{aligned} \tag{D.14}$$

Then, for any power m, *the expectation value* $\left\langle v^{\overline{n-1}}\right| (G^{\overline{n-1}})^m \left|v^{\overline{n-1}}\right\rangle$ *can be expressed in terms of the expectation values* $\left\langle v^{\bar{n}}\right| (G^{\bar{n}})^p \left|v^{\bar{n}}\right\rangle = \langle (G^{\bar{n}})^p\rangle$ *with* p *being at most* $m+2$. *The corresponding statement involving* H*'s and* $|w\rangle$*'s also holds.*

Proof. The first step is to prove that for any power m:

$$(G^{\overline{n-1}})^m = (G^{\bar{n}})^m + \sum_{i,j=0}^{m+1} \alpha_{i,j} (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j \tag{D.15}$$

Note that some of the $\alpha_{i,j}$ can be zero.

Indeed, we can use induction to prove Equation (D.15).
The base of the induction $m = 1$ gives us Equation (D.14), which holds.
Then, the power $m+1$ is:

$$(G^{\overline{n-1}})^{m+1} = (G^{\overline{n-1}})^m \cdot G^{\overline{n-1}}, \tag{D.16}$$

and substituting from Equation (D.14) and Equation (D.15), we get

$$\begin{aligned}
&(G^{\overline{n-1}})^{m+1} \\
&= \left[(G^{\bar{n}})^m + \sum_{i,j=0}^{m+1} \alpha_{i,j} (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j \right] \\
&\quad \cdot \left[G^{\bar{n}} + \frac{\langle (G^{\bar{n}})^3\rangle}{\langle (G^{\bar{n}})^2\rangle^2} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} - \frac{1}{\langle (G^{\bar{n}})^2\rangle} \left(G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^2 + (G^{\bar{n}})^2 \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| G^{\bar{n}} \right) \right] \\
&= (G^{\bar{n}})^{m+1} + \sum_{i,j=0}^{m+2} \alpha_{i,j} (G^{\bar{n}^n})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j,
\end{aligned}$$

which proves that Equation (D.15) holds for all m.
With this in place, we can proceed to prove our main claim about the corresponding expectation value:

$$\begin{aligned}
&\left\langle v^{\overline{n-1}}\right| (G^{\overline{n-1}})^m \left|v^{\overline{n-1}}\right\rangle \\
&= \left(\left\langle v^{\bar{n}}\right| - \frac{\langle G^{\bar{n}}\rangle}{\langle (G^{\bar{n}^n})^2\rangle} \left\langle v^{\bar{n}}\right| G^{\bar{n}} \right) \left((G^{\bar{n}})^m + \sum_{i,j=0}^{m+1} \alpha_{i,j} (G^{\bar{n}})^i \left|v^{\bar{n}}\right\rangle \left\langle v^{\bar{n}}\right| (G^{\bar{n}})^j \right) \left(\left|v^{\bar{n}}\right\rangle - \frac{\langle G^{\bar{n}}\rangle}{\langle (G^{\bar{n}})^2\rangle} G^{\bar{n}} \left|v^{\bar{n}}\right\rangle \right) \\
&= \langle (G^{\bar{n}})^m\rangle + a\langle (G^{\bar{n}})^{m+1}\rangle + b\langle (G^{\bar{n}})^{m+2}\rangle + \sum_{i,j=0}^{m+2} \alpha_{i,j} \langle (G^{\bar{n}})^i\rangle \langle (G^{\bar{n}})^j\rangle \\
&= \sum_{i,j=0}^{m+2} \alpha'_{i,j} \langle (G^{\bar{n}})^i\rangle \langle (G^{\bar{n}})^j\rangle,
\end{aligned}$$

which completes our proof that the highest power is $m+2$ for any m. Notice that we did not fully specified the scalar factors $a, b, \alpha_{i,j}, \alpha'_{i,j}$, as it is easy to verify that they do not contain any higher powers (as in the previous lemma). □

Bibliography

[1] N Aharon and J Silman. Quantum dice rolling: a multi-outcome generalization of quantum coin flipping. *New Journal of Physics*, 12(3):033027, mar 2010.

[2] Nati Aharon, André Chailloux, Iordanis Kerenidis, Serge Massar, Stefano Pironio, and Jonathan Silman. Weak coin flipping in a device-independent setting. In *Revised Selected Papers of the 6th Conference on Theory of Quantum Computation, Communication, and Cryptography - Volume 6745*, TQC 2011, pages 1–12, New York, NY, USA, 2014. Springer-Verlag New York, Inc.

[3] Dorit Aharonov, André Chailloux, Maor Ganz, Iordanis Kerenidis, and Loïck Magnin. A simpler proof of existence of quantum weak coin flipping with arbitrarily small bias. *SIAM Journal on Computing*, 45(3):633–679, jan 2014.

[4] Andris Ambainis. A new protocol and lower bounds for quantum coin flipping. *Journal of Computer and System Sciences*, 68(2):398–416, 2004.

[5] Atul Singh Arora, Jérémie Roland, and Stephan Weis. Weak Coin Flipping, 2018.

[6] C. H. Bennett and G. Brassard. Public-key distribution and coin tossing. In *Int. Conf. on Computers, Systems and Signal Processing*, pages 175–179, 1984.

[7] Guido Berlín, Gilles Brassard, Félix Bussières, and Nicolas Godbout. Fair loss-tolerant quantum coin flipping. *Physical Review A*, 80(6), dec 2009.

[8] Rajendra Bhatia. *Matrix Analysis*. Springer New York, 2013.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, mar 2004.

[10] AndrÃ© Chailloux and Iordanis Kerenidis. Optimal Bounds for Quantum Bit Commitment. In *52nd FOCS*, pages 354–362, 2011.

[11] André Chailloux, Gus Gutoski, and Jamie Sikora. Optimal bounds for semi-honest quantum oblivious transfer. 2013.

[12] André Chailloux and Iordanis Kerenidis. Optimal Quantum Strong Coin Flipping. In *50th FOCS*, pages 527–533, 2009.

[13] Claude Crépeau. Quantum oblivious transfer. *Journal of Modern Optics*, 41(12):2445–2454, dec 1994.

[14] Tobias Fritz. Does the set of operator monotone functions become larger if we restrict ourselves to real symmetric matrices? Posted on MathOverflow as a comment (sister site of StackOverflow)., 2018.

[15] Maor Ganz. Quantum Leader Election. 2009.

[16] Oded Goldreich, , Silvio Micali, Avi Wigderson, and and. How to play any mental game, or a completeness theorem for protocols with honest majority. In *Providing Sound Foundations for Cryptography: On the Work of Shafi Goldwasser and Silvio Micali*. Association for Computing Machinery, oct 2019.

[17] Lov K. Grover. A fast quantum mechanical algorithm for database search. In *28th STOC*, pages 212–219, 1996.

[18] Gus Gutoski, Ansis Rosmanis, and Jamie Sikora. Fidelity of quantum strategies with applications to cryptography. *Quantum*, 2:89, sep 2018.

[19] William W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, jun 1989.

[20] Thomas Van Himbeeck, Erik Woodhead, Nicolas J. Cerf, Raúl García-Patrón, and Stefano Pironio. Semi-device-independent framework based on natural physical assumptions. *Quantum*, 1:33, nov 2017.

[21] Peter Høyer and Edouard Pelchat. Point Games in Quantum Weak Coin Flipping Protocols. Master's thesis, University of Calgary, 2013.

[22] I. Kerenidis and A. Nayak. Weak coin flipping with small bias. *Information Processing Letters*, 89(3):131–135, feb 2004.

[23] A. Kitaev. Quantum coin flipping. Talk at the 6th workshop on Quantum Information Processing, 2003.

[24] Hoi-Kwong Lo and H.F. Chau. Why quantum bit commitment and ideal quantum coin tossing are impossible. *Physica D: Nonlinear Phenomena*, 120(1):177 – 187, 1998. Proceedings of the Fourth Workshop on Physics and Consumption.

[25] Dominic Mayers. Unconditionally secure quantum bit commitment is impossible. *Physical Review Letters*, 78(17):3414–3417, apr 1997.

[26] Carl A. Miller. The impossibility of efficient quantum weak coin-flipping.

[27] Carlos Mochon. Large family of quantum weak coin-flipping protocols. *Phys. Rev. A*, 72:022341, 2005.

[28] Carlos Mochon. Quantum weak coin flipping with arbitrarily small bias. *arXiv:0711.4114*, 2007.

[29] Ashwin Nayak and Peter Shor. Bit-commitment-based quantum coin flipping. *Phys. Rev. A*, 67:012304, Jan 2003.

[30] Ashwin Nayak, Jamie Sikora, and Levent Tunçel. A search for quantum coin-flipping protocols using optimization techniques. *Mathematical Programming*, 156(1-2):581–613, may 2014.

[31] Ashwin Nayak, Jamie Sikora, and Levent Tunçel. Quantum and classical coin-flipping protocols based on bit-commitment and their point games. 2015.

[32] Michael O Rabin. How to exchange secrets with oblivious transfer. *IACR Cryptology ePrint Archive*, 2005:187, 2005.

[33] Rolf Schneider. *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge University Press, 2009.

[34] Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, mar 1950.

[35] P.W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*. IEEE Comput. Soc. Press, 1994.

[36] Jamie Sikora and John H. Selby. On the impossibility of coin-flipping in generalized probabilistic theories via discretizations of semi-infinite programs. 2019.

[37] R. W. Spekkens and Terry Rudolph. Quantum protocol for cheat-sensitive weak coin flipping. *Physical Review Letters*, 89(22), nov 2002.

[38] Stephen Wiesner. Conjugate coding. *ACM SIGACT News*, 15(1):78–88, jan 1983.

[39] Andrew Chi-Chih Yao. Security of quantum protocols against coherent measurements. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing - STOC '95*. ACM Press, 1995.