COPPE
UFRJ
Instituto Alberto Luiz Coimbra de
Pós-Graduação e Pesquisa de Engenharia

A STATISTICAL APPROACH TO ANALYZING THE QUANTUM ALTERNATING OPERATOR ANSATZ WITH GROVER MIXER

Guilherme Adamatti Bridi

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Franklin de Lima Marquezino

Rio de Janeiro
Março de 2024

A STATISTICAL APPROACH TO ANALYZING THE QUANTUM ALTERNATING OPERATOR ANSATZ WITH GROVER MIXER

Guilherme Adamatti Bridi

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Orientador: Franklin de Lima Marquezino

Aprovada por: Prof. Franklin de Lima Marquezino
Prof. Renato Portugal
Prof. Fabio Pereira dos Santos

RIO DE JANEIRO, RJ – BRASIL
MARÇO DE 2024

Bridi, Guilherme Adamatti

A Statistical Approach to Analyzing the Quantum Alternating Operator Ansatz with Grover Mixer/Guilherme Adamatti Bridi. – Rio de Janeiro: UFRJ/COPPE, 2024.

XVIII, 157 p.: il.; 29, 7cm.

Orientador: Franklin de Lima Marquezino

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2024.

Referências Bibliográficas: p. 139 – 149.

1. QAOA. 2. Grover mixer. 3. GM-QAOA. 4. GM-Th-QAOA. 5. Combinatorial optimization. I. Marquezino, Franklin de Lima. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

"Every kid starts out as a natural-born scientist, and then we beat it out of them. A few trickle through the system with their wonder and enthusiasm for science intact." (Carl Sagan)

Acknowledgments

Primeiramente, agradeço à minha mãe, Noeli, a pessoa mais importante da minha vida. Reconheço que tenho dificuldades em demonstrar e principalmente em falar meus sentimentos, mas quero que saiba que meu amor é genuíno.

Agradeço ao meu avô Iter (*in memoriam*) e à minha avó Nadir por serem importantes na minha criação durante boa parte da minha infância e adolescência. Uma parte vital da minha memória afetiva corresponde aos verões que passamos juntos na praia de Arroio do Sal. Em especial, presto uma pequena homenagem ao meu avô, que deixou como legado um exemplo ímpar de ser humano.

Agradeço ao meu pai, Carlos (*in memoriam*), que, apesar de não termos convivido de maneira consistente, guardo ótimas lembranças.

Agradeço à minha tia Sirlei, que sempre esteve presente em minha vida.

Agradeço à minha prima Bruna, minha referência de pessoa inteligente desde a infância.

Agradeço à Santina Bado, por ter me cuidado durante grande parte da minha infância. Também devo mencionar sua gentileza em me acolher em sua casa enquanto eu realizava estágio. Por motivos similares a este último, agradeço à minha tia-avó Maria.

Agradeço aos meus amigos da graduação, Alex Barbosa, Arthur Rossato e Aragones Gonçalves, cuja amizade levo para vida. Um agradecimento especial ao Alex por sua generosidade ao me emprestar seu notebook em um momento crucial do mestrado.

Agradeço ao meu amigo, Oscar Martins, pela parceria e apoio mútuo tanto nos momentos bons quanto nos momentos difíceis ao longo do mestrado.

Agradeço ao meu amigo de república Gustavo Bezerra pela companhia, parceria e a oportunidade de morar em um local mais pacato.

Agradeço aos professores André Michel, André Martinotto e Alexandre Mesquita pela grande ajuda na busca de um programa de mestrado na área de computação quântica. Em particular, agradeço ao Alexandre, meu orientador de TCC na graduação, não apenas pela orientação, mas por casualmente ter sugerido a área de computação quântica como tema de pesquisa, um ponto de contingência na intrincada e imensuravelmente complexidade teia de eventos da vida, responsável por transformar minha trajetória acadêmica.

Agradeço aos membros avaliadores da banca, Renato Portugal e Fabio Santos, por aceitarem participar deste processo. Em particular, agradeço ao Renato por permitir que eu realizasse seminários para seu grupo de pesquisa, o que enriqueceu minha experiência acadêmica e contribuiu diretamente para esta dissertação; e ao Fábio Santos pela disponibilização do *cluster* para realização de experimentos numéricos fundamentais em dado momento da pesquisa.

Agradeço ao meu orientador, Franklin Marquezino, não apenas pela excelente orientação ao longo do mestrado, mas também pela receptividade desde o momento em que eu, uma pessoa desconhecida, estava buscando um programa de pós-graduação. É seguro dizer que, sem seu apoio, esta dissertação não seria possível.

Agradeço à CAPES pelo suporte financeiro.

Por fim, agradeço a todas as pessoas não mencionadas que direta ou indiretamente contribuíram para este trabalho.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA ABORDAGEM ESTATÍSTICA PARA ANALISAR O QUANTUM ALTERNATING OPERATOR ANSATZ COM GROVER MIXER

Guilherme Adamatti Bridi

Março/2024

Orientador: Franklin de Lima Marquezino

Programa: Engenharia de Sistemas e Computação

O operador Grover mixer é uma versão variacional do operador difusão de Grover, introduzido como um operador mixing para o *Quantum Alternating Operator Ansatz* (QAOA) e usado em duas variantes, conhecidas como Grover Mixer QAOA (GM-QAOA) e Grover Mixer Threshold QAOA (GM-Th-QAOA). Uma propriedade importante dessas variantes é que o valor esperado é invariante para qualquer permutação de estados. Como consequência, o algoritmo é independente da estrutura do problema. Se, por um lado, esta característica levanta sérias dúvidas sobre a capacidade do algoritmo de superar o *bound* do problema de busca não estruturada, por outro lado, pode abrir caminho para o seu estudo analítico. Neste sentido, este trabalho considera uma abordagem estatística para analisar tanto o GM-QAOA quanto o GM-Th-QAOA que resulta em expressões analíticas para o valor esperado que dependem da distribuição de probabilidade associada ao espectro do hamiltoniano do problema. Embora no caso do GM-QAOA a expressão dependa exponencialmente do número de camadas, o caso mais simples do GM-Th-QAOA resulta em uma expressão independente desse parâmetro e, com ela, fornecemos limites para diferentes métricas de desempenho. Posteriormente, estendemos a análise do GM-Th-QAOA para um contexto mais geral para o QAOA com o Grover mixer que chamamos de *Grover-based QAOA*. Nessa estrutura, que permite ao operador separador de fase codificar qualquer compilação da função de custos, generalizamos todos os *bounds* utilizando um argumento de contradição com a otimalidade do algoritmo de Grover no problema de busca não estruturada. Como resultado, obtemos a principal contribuição deste trabalho, formalizando a noção de que o Grover mixer reflete, no máximo, uma aceleração quadrática ao estilo do algoritmo de Grover sobre a força bruta clássica.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A STATISTICAL APPROACH TO ANALYZING THE QUANTUM ALTERNATING OPERATOR ANSATZ WITH GROVER MIXER

Guilherme Adamatti Bridi

March/2024

Advisor: Franklin de Lima Marquezino

Department: Systems Engineering and Computer Science

The Grover mixer operator is a variational version of Grover's diffusion operator, introduced as a mixing operator for the Quantum Alternating Operator Ansatz (QAOA) and used in two variants known as Grover Mixer QAOA (GM-QAOA) and Grover Mixer Threshold QAOA (GM-Th-QAOA). An important property of these variants is that the expectation value is invariant over any permutation of states. As a consequence, the algorithm is independent of the structure of the problem. If, on the one hand, this characteristic raises serious doubts about the capacity of the algorithm to overcome the bound of the unstructured search problem, on the other hand, it can pave the way to its analytical study. In this sense, this work considers a statistical approach to analyze both GM-QAOA and GM-Th-QAOA that results in analytical expressions for the expectation value depending on the probability distribution associated with the problem Hamiltonian spectrum. Although in the case of GM-QAOA, the expression depends exponentially on the number of layers, the more simple case of GM-Th-QAOA results in an expression with complexity independent of that parameter, and, with it, we provide bounds for different performance metrics. Subsequently, we extend the analysis of GM-Th-QAOA to a more general context for QAOA with the Grover mixer we called Grover-based QAOA. In that framework, which allows the phase separation operator to encode any compilation of the cost function, we generalize all the bounds by using a contradiction argument with the optimality of Grover's algorithm on the unstructured search problem. As a result, we get the main contribution of this work, formalizing the notion that the Grover mixer, at most, reflects a quadratic Grover-style speed-up over classical brute force.

Contents

List of Figures **xii**

List of Abbreviations **xvii**

1 Introduction **1**

1.1 Contributions 4

1.2 Roadmap and general comments 6

2 A Review on Probability Theory **8**

2.1 Probability spaces 8

2.1.1 Properties of probability 10

2.1.2 Conditional probability 10

2.1.3 Independence of events 11

2.2 Random variables 11

2.2.1 Cumulative distribution function 12

2.2.2 Probability distribution 13

2.2.3 Quantile function 16

2.2.4 Random vectors 17

2.2.5 Independence of random variables 17

2.2.6 Sum of independent random variables 18

2.3 Families of probability distributions 18

2.4 Moments 21

2.4.1 Expectation 21

2.4.2 Variance and standard deviation 22

2.4.3 Generalizing the moments 23

2.5 Random sample 28

2.5.1 Estimators and confidence intervals 28

2.5.2 Central limit theorem 29

2.6 Probability bounds 30

2.6.1 Markov and Chebyshev inequalities 30

2.6.2 Jensen's inequality 31

2.7 Conditional probability distribution 31
2.7.1 Truncated distribution 32
2.8 Characteristic function 34

3 The Quantum Alternating Operator Ansatz and the Grover Mixer 37
3.1 Combinatorial optimization 38
3.1.1 Combinatorial optimization problems 38
3.1.2 Decisions problems and complexity classes 40
3.1.3 Classical approaches 41
3.2 Grover's algorithm and the unstructured search problem 43
3.2.1 Geometric interpretation of Grover's algorithm 45
3.2.2 The low-convergence regime and the maximum amplification . 46
3.2.3 Optimality of Grover's algorithm on the unstructured search problem 46
3.3 Grover Adaptive Search 48
3.4 Variational Quantum Algorithms 49
3.4.1 Ansatzes 50
3.4.2 Optimizers and the barren plateau phenomena 51
3.4.3 Applications 52
3.5 Quantum Adiabatic Algorithm 52
3.6 Quantum Approximate Optimization Algorithm 54
3.7 Quantum Alternating Operator Ansatz 55
3.7.1 Phase separation and mixing operators 56
3.7.2 Angles finding and analytical results 58
3.8 Grover mixer 59
3.8.1 Grover Mixer Quantum Alternating Operator Ansatz 59
3.8.2 Grover Mixer Threshold Quantum Alternating Operator Ansatz 60
3.8.3 Classical simulation of QAOA with Groxer mixer 63
3.9 Quantum Walk-based Optimization Algorithm 64
3.10 Maximum Amplification Optimization Algorithm 65
3.11 Performance of QAOA with Grover mixer 66

4 The Statistical Approach 69
4.1 GM-QAOA analysis 72
4.1.1 Depth 1 73
4.1.2 Arbitrary depth 76
4.1.3 Upper bounds on the standard score for GM-QAOA 83
4.1.4 The binary function 84
4.2 GM-Th-QAOA analysis 88
4.2.1 Threshold curve problem 92

4.2.2 Asymptotic tight bound on quantile 93

4.2.3 Upper bounds on the standard score for GM-Th-QAOA 95

4.2.4 Combining the bounds on the standard score and quantile . . 99

5 Numerical experiments 103

5.1 Normal distribution 103

5.1.1 Single layer 104

5.1.2 Discretization 105

5.1.3 Scaling the number of layers 108

5.2 Outliers effect 110

5.3 Asymptotic scale of the quadratic speed-up 113

5.4 Near optimal asymptotic standard score 114

5.5 Discrete distributions 115

5.5.1 Binomial distribution 116

5.5.2 Uniform distributions 118

6 Bounds on Grover-based Quantum Alternating Operator Ansatz 121

6.1 The Grover-based Quantum Alternating Operator Ansatz 121

6.2 General bounds 123

6.3 Bounds on Max-Cut 131

7 Conclusion 135

References 139

A Low-convergence Regime Gives the Maximum Amplification on Grover's Search 150

B Proof of Theorem 14 152

List of Figures

2.1 Probability density functions of the random variable X given by the distribution Normal(u, s^2) with $u = 3$ and $s = 2$, and its associated Y and Z random variables. The random variable Y is a location shifting to get zero expectation, while Z is additionally a change of the scale to get a unit standard deviation. 25

2.2 Graphic interpretation of skewness. (a) The right-skewed distribution Chi-squared(k) with 4 degrees of freedom. The skewness is $\sqrt{2}$, the mean 4, and the median ≈ 3.5679. (b) The left-skewed RChi-squared(k) with 4 degrees of freedom. The skewness is $-\sqrt{2}$, the mean -4, and the median ≈ -3.5679. These specific distributions follow the described "rule" that relates skewness and mean/median positions. 26

2.3 Log-linear graphic of the pdf of the standardized distributions $f_Z(x)$ of continuous uniform, normal, logistic, and Laplace distributions. All of them are symmetric around the mean distributions without a shape parameter. The choice of parameters to get Z gives CUniform$(-\sqrt{3}, \sqrt{3})$, Normal$(0, 1)$, Logistic$(0, \sqrt{3}/\pi)$, and Laplace$(0, \sqrt{2}/2)$, with the kurtosis given respectively by 9/5, 3, 21/5, and 6. The continuous uniform distribution, since is zero for $|x| > \sqrt{3}$, has a smaller impact on outliers and presents the lowest kurtosis between the distributions. To the remainder distributions, from a certain value of $|x|$, the pdf of normal distribution decays more quickly than logistic distribution, which in turn decays faster than Laplace distribution. Therefore, the increasing order of the impact of outliers, given by normal, logistic, and Laplace distributions, fits with the increasing order of their kurtosis values. 27

3.1 Geometric interpretation of a single Grover's iteration. Two-dimensional subspace spanned by the vectors $|b\rangle$ and $|a\rangle$. The initial state $|d\rangle$ makes an angle of $\theta/2$ with the axis $|a\rangle$. The application of the oracle and diffusion operator are reflections over $|a\rangle$ and $|d\rangle$, respectively, which results in an angle of $3\theta/2$ for $G|d\rangle$. The creation of this figure by the author was inspired by Figure 6.3 of Nielsen and Chuang [1]. 45

3.2 Amplification η versus the ratio ρ for the range 10^{-5} until 0.25 on linear-log scale and $r = 1$. The ratio $r = 0.25$, which has an amplification of $\eta = 4$, gives the $\pi/2$ radians of the geometric interpretation of Grover's algorithm. We normalize η by the maximum amplification of $(2r+1)^2$. For low ratios, the amplification gets closer to the maximum amplification, while for ratios near to $\pi/2$ radians on geometric interpretation, η is considerably lower than $(2r+1)^2$. 47

4.1 (a) Graph of an instance of Max-Cut problem with 5 vertices and $M = 32$. (b) Probability mass function $f_X(x)$ associated with the instance of (a). We multiply the solutions by -1 to convert it into a minimization problem. The optimal solution has cost $R_X^{min} = -6$ and is given by the partition of the vertices into $\{1,3,4\}$ and $\{2,5\}$. The maximum solution is given by $R_X^{max} = 0$, with trivial partition into $\{1,2,3,4,5\}$ and $\varnothing$. Note that all partitions are duplicates on the binary codification of the problem. 71

4.2 (a) The inclination of the curve $C^{Th}(r)$ versus r asymptotically converges to a certain value. (b) The ratio ρ that maximizes $C_r(t)$ scales with $1/r^2$, with a constant that asymptotically converges also to a certain value. 98

5.1 Threshold curve for $r = 1$ on normal distribution. Without loss of generality, we consider $C_r(t)$ versus T/σ as the threshold curve instead of the original $E_r(t)$ versus T of the Subsec. 4.2.1. T/σ means the normalized threshold for any choice of u and s. The resolution used is 2000 values of threshold, chosen uniformly between the interval of -4 and 3. 104

5.2 The landscape of GM-QAOA for $r = 1$ on the normal distribution. The range considered is $\gamma, \beta \in (\pi, \pi]$. We observe two symmetric global minima and two symmetric global maxima in the same way as discussed at the end of Subsec. 4.1.1. 106

5.3 Simulation of distribution Normal(u, s^2) for GM-Th-QAOA and CRS up to 10^6 layers, and GM-QAOA up to 8 layers. (a) Standard score generically denoted by C versus r in a linear-log scale graphic. (b) Log-log graphic of the quantile achieved by the algorithms, generically denoted $F_X(\mathrm{E})$, as a function of r. 109

5.4 Optimal threshold on the simulation of distribution Normal(u, s^2) for GM-Th-QAOA up to 10^6 layers. (a) Linear-log graphic of $|T|/\sigma$ versus r, compared with $C_r(t)$. The absolute value of T is taken to compare directly with $C_r(t)$. (b) Log-log graphic of $F_X(t)$ versus r, compared with $F_X(E_r(t))$. (c) For that optimal threshold value, the figure shows the probability $P(\rho, r)$ on a graphic with a log-log scale for $1 - P(\rho, r)$ versus r. 110

5.5 Decay of the cdf of the standard reflected exponential distribution, compared to the standard normal, standard logistic, and standard Laplace distributions. The scale of both graphics is log-linear. (a) Larger range of $x \in [-7, 2]$. (b) Zoom on the range $x \in [-3, 0.5]$. . . . 111

5.6 Simulation of normal, logistic, Laplace, and reflected exponential distributions for GM-QAOA up to 8 layers. (a) $C_r(\boldsymbol{\beta}, \boldsymbol{\gamma})$ versus r. (b) $F_X(E_r(\boldsymbol{\beta}, \boldsymbol{\gamma}))$ versus r in log-log scale. 112

5.7 Simulation of normal, logistic, Laplace, and reflected exponential distributions for GM-Th-QAOA up to 10^6 layers. (a) $C_r(t)$ versus r on linear-log scale. (b) $F_X(E_r(t))$ versus r in log-log scale. 112

5.8 Distribution RGamma(a, b) with $b = 1/2$ and $a = k/2$ for values of $k = 2$ and then decreasing with powers of 10 such that $k = 10^{-j}$ for $j = 2, 4, 6, 8$. Note that if k were a positive integer, we would have the chi-squared distribution of k degrees of freedom and that the case of $a = 2$ and $b = 1/2$ reduce the reflected distribution gamma to the reflected exponential distribution with $l = 1/2$. (a) $F_X(\mu)$ versus k on log-log scale from 10^{-8} to 2 with 10000 values of resolution. (b) Log-log graphic of $F_X(E_r(t))$ versus r up to 10^5 rounds. 113

5.9 Standard score achieved by GM-Th-QAOA up to 10^5 rounds for Pareto(ϵ, x_m) with different values of j. For viewing purposes, we normalize $C_r(t)$ by C_{max}, where C_{max} is the value of $C_{10^5}(t)$. 115

5.10 Simulation of Binomial(n,p) with $n = 200$ and $p = 0.5$ for GM-Th-QAOA up to 100 rounds, compared with the distribution Normal(u, s^2). (a) $C_r(t)$ versus r on the linear-log scale and (b) $F_X(E_r(t))$ versus r on log-log scale. (b) $F_X(t)$ versus r on log-log scale. 116

5.11 (a) The optimal threshold of GM-Th-QAOA and its probability (b) $P(\rho, r)$ (we also show $P(\rho, r)$ of the normal distribution as background) versus r in linear-log scale for the distribution Binomial(n, p) with $n = 200$ and $p = 0.5$ up to 100 rounds. By plot (a), the threshold value starts to stagnate for some rounds after a certain point. For a given value of optimal threshold t, evolving the number of rounds, the probability $P(\rho, r)$ increases until eventually arriving at the maximum value of 1, as observed in the plot of (b). From there, the only way to improve the performance of GM-Th-QAOA is by changing the threshold to the next value, $t - 1$. However, we may need more than one round for the change to be advantageous, and thus, the algorithm stagnates in that interval. Upon reaching $t - 1$, probability returns to below 1, and the process repeats, which explains the behavior of Fig. 5.10. Indeed, we can observe that the points with probability 1 of the plot (b) match the stagnation points of Fig. 5.10. 117

5.12 Threshold curve of distributions (a) Normal(u, s^2) and (b) Binomial(n, p) ($n = 200$ and $p = 0.5$) with $C_r(t)$ versus $F_X(t)$ on a linear-log scale. The resolution considered on the continuous distribution was of 2000 values for the threshold. For viewing purposes, we show only the values of r in terms of powers of 10 from 1 up to 10^6. The envelope that unites the curves is the interval of $P(\rho, r) = 1$. 118

5.13 The standard score achieved by GM-QAOA and GM-Th-QAOA, generically denoted by C, versus r, up to 8 rounds for the distribution CUniform(a, b). We indicate the maximum achievable standard score $\sqrt{3}$ with a line segment. 119

5.14 Standard score achieved by GM-Th-QAOA up to 20 layers for the distributions CUniform(a, b) and DUniform$(-n, n)$, with n increasing with powers of 2 such as 2^j for $j = 1, 2, 3, 4, 5, 6$. The graphic is a log-linear of $\sqrt{3} - C_r(t)$ versus r. 120

6.1 (a) Probability mass function and (b) cumulative distribution function concerning the random variable Y for the Max-Cut problem on the graph $K_{50,50}$. We compute efficiently the explicit distribution from our characterization of the solution space given by Eq. (6.39). Both graphs are on a log-linear scale. As the possible values for the cost are given by $\frac{1}{2}(n - 2j)(n - 2k)$ for all $0 \leq j, k \leq n$, the region near to 0 is denser than the regions near to $R_{X_c}^{min}$ and R_{X}^{max}. Although the decay is not uniform, the general trend is clearly exponential. 133

6.2 (a) Linear-log base 2 plot of approximation ratio versus the number of layers considering the application of the maximum amplification bound for the graph $K_{50,50}$ on Max-Cut. As expected, the growth rate is logarithmic. The resolution used on the number of layers (due to the extremely high number of layers to get $\lambda = 1$) is $\lceil 2^{x/100} \rceil$ for $x = 0, 1, \ldots, 5000$. Since the probability distribution is discrete, the values τ_1 and τ_2 on Eq. (6.21) were obtained via testing the point of the support via brute force. (b) The minimum number of layers required to maximum amplification bound achieves three different values of approximation ratio on the Max-Cut with the graph $K_{n,n}$ for $n = 4, 5, \ldots, 100$. The considered values of λ are $\lambda = 1$, in which we calculate analytically that $r = \lceil 2^{0.5(|\mathcal{V}|-3)} \rceil$; $\lambda = 16/17 \approx 0.9412$, the λ value in which Max-Cut becomes NP-Hard; and the approximation ratio guaranteed by the classical Goemans-Williamson algorithm, given by $\lambda \approx 0.8786$. The scale is log-linear with base 2. The value of r in which the approximation ratio is achieved was efficiently found with a binary search. As predicted, the number of layers scales exponentially in all of them. (c) The same as (b), but with $n = 4, 5, \ldots, 300$ and $\lambda = 0.52$. As the expectation value of a uniform superposition gives $\lambda = 0.5$, this approximation ratio is extremely low. However, even for such low performance, given a sufficient number of vertices, we observe the exponential dependence on the number of layers r. . . 134

List of Abbreviations

AQC	Adiabatic Quantum Computation, p. 52
BFGS	Broyden Fletcher, Goldfarb, Shanno (BFGS), p. 58
BP	Barren plateau, p. 51
CF	Characteristic function, p. 34
CLT	Central limit theorem, p. 29
COP	Combinatorial optimization problems, p. 2
CPBO	Constrained Polynomial Binary Optimization, p. 49
CRS	Classical random sampling, p. 42
CTQW	Continuous-time Quantum Walk, p. 2
DGK	Discrete Gaussian kernel, p. 20
GAS	Grover Adaptive Search, p. 6
GM-QAOA	Grover Mixer Quantum Alternating Operator Ansatz, p. 2
GM-Th-QAOA	Grover Mixer Threshold Quantum Alternating Operator Ansatz, p. 2
HAS	Hesitant Adaptive Search, p. 48
HHL	Harrow–Hassidim–Lloyd, p. 48
LOTUS	Law of the unconscious statistician, p. 22
MGF	Moment generating functions, p. 34
NISQ	Noisy Intermediate-Scale Quantum, p. 1
QAA	Quantum Adiabatic Algorithm, p. 1

QAOA	Quantum Approximate Optimization Algorithm or Quantum Alternating Operator Ansatz, depending on the context, p. 1
QFT	Quantum Fourier Transform, p. 64
QUBO	Quadratic Unconstrained Binary Optimization, p. 49
QWOA	Quantum Walk-based Optimization Algorithm, p. 2
RGAS	Restricted Grover Adaptive Search, p. 66
SGD	Stochastic Gradient Descent, p. 51
Th-QAOA	Threshold Quantum Alternating Operator Ansatz, p. 2
UCC	Unitary Coupled Clustered, p. 50
VQA	Variational Quantum Algorithms, p. 1
VQE	Variational Quantum Eigensolver, p. 52
VQF	Variational Quantum Factoring, p. 52
VQLS	Variational Quantum Linear Solver, p. 52
cdf	Cumulative distribution function, p. 12
i.i.d.	Independent and identically distributed, p. 18
iCANS	Individual Coupled Adaptive Number of Shots, p. 51
pdf	Probability density function, p. 13
pmf	Probability mass function, p. 13

Chapter 1

Introduction

The current state-of-the-art quantum computing technology is known as the Noisy Intermediate-Scale Quantum (NISQ) era [2, 3]. During the NISQ era, we have access to quantum computers with only tens or hundreds of qubits. As challenging as the limitation on the number of qubits, is the presence of noise, which results in coherent and incoherent errors that compromise the quality of output measurements and consequently restrict the allowed depth of quantum circuits [4]. Unfortunately, the protocols of quantum error correction, due to the large qubit requirements, are far beyond NISQ device capabilities, in such a way that if we are interested in achieving a speed-up of quantum algorithms over classical ones—often called quantum advantage—at scales that are of practical interest in the short or medium terms, we need to handle the limitations of NISQ devices [4].

In this context, the Variational Quantum Algorithms (VQA) [4], a class of hybrid quantum-classical algorithms, have gained prominence in recent years as a potentiality of quantum advantage in the NISQ era. These algorithms use the power of classical computing to help overcome current technology limitations of quantum computing. Specifically, they work with parameterized quantum circuits of limited depth and number of qubits, using classical optimizers that use strategies based on optimization or training to iteratively update the optimization parameters aiming extremize[1] a function based on observables measured from the quantum circuit. There are several algorithms and classes of algorithms in the VQAs context, encompassing a wide range of tasks, from quantum chemistry and quantum physics to combinatorial optimization and mathematical applications.

One of the most prominent cases of VQA is the Quantum Approximate Optimization Algorithm (QAOA) [5], which can be generalized to the Quantum Alternating Operator Ansatz (QAOA) [6]. QAOA is a class of algorithms derived from the Quantum Adiabatic Algorithm (QAA) [7, 8] and used heuristically to find

[1] Minimize or maximize.

solutions to combinatorial optimization problems (COP) [9]. An extensive list of optimization problems has already been considered in the context of QAOA, being likely the Max-Cut problem the most well-studied [10]. The algorithm consists of a given number of rounds of alternating application of two parameterized operators in an initial state. The first is the phase separation operator, which changes the relative phases between states according to the cost function (the objective function of the COP), while the last one is the mixing operator, responsible for generating interference between the states, changing its amplitudes [11, 12].

The original mixing operator of QAOA uses the transverse field mixer Hamiltonian [5], which is given by a sum of Pauli-X operators. Since then, many other variations with different types of mixers have already been introduced in the literature [6, 12–15]. There is numerical evidence that the choice of mixing operator significantly affects the performance of QAOA [11, 16, 17] and therefore choosing the ideal mixer for a given optimization problem is an important research topic. One variant of particular interest is the Grover Mixer Quantum Alternating Operator Ansatz (GM-QAOA) [13]. The mixing operator of GM-QAOA is a variational version of Grover's diffusion operator of Grover's algorithm [18, 19] called Grover mixer operator. In GM-QAOA formulation, a necessary condition to the construction of the mixing operator is the existence of an efficient[2] preparation of uniform superposition over the feasible states—which covers problems like the Traveling Salesman Problem, the Max k-Vertex Cover, and the Discrete Portfolio Rebalancing. Alternatively, the operator of Grover mixer can be constructed with the formulation of Quantum Walk-based Optimization Algorithm (QWOA) [12, 14], a generalization of QAOA that interprets the mixing operator as a Continuous-time Quantum Walk (CTQW) operator [20–22]. In that case, the Grover mixer operator is equivalent to QWOA on the complete graph up to a change of the scale on the operator parameter [23]. Problems such as the Capacitated Vehicle Routing [24] and the Portfolio Optimization [25] have already been numerically studied within the QWOA framework.

Another variant of Quantum Alternating Operator Ansatz using the Grover mixer is the so-called Grover Mixer Threshold QAOA (GM-Th-QAOA) [26], an algorithm combing such mixing operator with the more general Threshold QAOA (Th-QAOA), which in turn changes the original phase separation to encode a compilation of the cost function into a threshold function splitting the solution space from a value. In particular, the choice of all angles as being equal to π reduces the GM-Th-QAOA to Grover's algorithm for marked states above (considering the original

[2] In this work, we use the terminology efficient for its well-known meaning in the complexity theory, i.e., an efficient algorithm means an algorithm with runtime upper bounded by a polynomial function on the size of the input of the problem.

definition) a given threshold. An advantage of this variant is admitting an efficient procedure to find optimal parameters—the angles and the threshold value—that eliminates the costly variational loop of the usual QAOA. Furthermore, it has been numerically observed that the performance of GM-Th-QAOA consistently overcomes GM-QAOA in all instances considered [11, 17, 26].

The performance of QAOA with the Grover mixer, individually or compared with other mixers, has already been considered in the literature. The initial thought, corroborated by numerical experiments on small instances, was that the Grover mixer would overcome transverse field mixer due to its ability to mix quickly and its global symmetry among states [12, 16, 25]. However, later experiments on larger instances indicated that this advantage soon disappeared with Grover mixer losing to transverse field mixer [17] and performing even exponentially worse [11] than the clique mixer [27, 28]. One can argue, summarizing insights and conclusions from these recent studies, that the worst performance of the Grover mixer may be due to the fact that it depends only on the distribution of the solution space and the algorithm does not see the structure of the optimization problem and possibly is limited to the bound of the unstructured search problem [29], drastically compromising algorithm performance on large instances. On the other hand, other mixers, such as transverse field and clique, could, in principle, overcome that limit by exploiting the underlying problem structure.

Despite the performance limitation, the Grover mixer provides a unique opportunity to get analytical studies for QAOA. Historically, analytical results for QAOA are rare and sparse due to the high complexity of the algorithm [11]. However, the independence of the structure of the Grover mixer can greatly simplify the analysis. That has been noticed by Bennett and Wang [23], who used degeneracy in solution space to make edge contractions on the complete graph of QWOA. Headley and Wilhelm [30] went further and introduced a statistical approach (i.e., an approach based on Probability Theory [31, 32]) using random variables to model the problem that led to an analytical expression of the expectation value of GM-QAOA depending only on the probability distribution associated with the problem Hamiltonian spectrum—the solution space of the optimization problem. The prominent statistical quantity of the resulting expression is the characteristic function, i.e., the Fourier transform of the probability distribution. Although the complexity of that expression scales exponentially with $\mathcal{O}(4^r)$, where r is the number of layers, it does not depend on the size of the problems, which allows computing the optimal parameter (or near-optimal) in size limit for problems with instances that converge asymptotically towards a fixed distribution, such as the Number Partition Problem with independent and identically distributed choice of the numbers. The statistical approach was later generalized by Headley [33] to the structure-dependent transverse field mixer and the so-called line mixer, in a work that establishes how QAOA can actually exploit the underlying structure of the problem.

1.1 Contributions

The main results of the present work can be divided into two parts. The first, present on Sec. 4.1, is equivalent to the aforementioned statistical approach of Headley and Wilhelm [30] for GM-QAOA. We develop it while working in the context of QWOA on the complete graph before becoming aware of the existence of Headley and Wilhelm [30] paper. As the results of the second part are a direct continuation of the first one, we decided to keep it in the text. Furthermore, some differences between our text and the work of Headley and Wilhelm [30] are listed.

- Headley and Wilhelm [30] use continuous random variables to model the solution space as an asymptotic approximation on the limit of large size. We use the exact case of discrete random variables, assuming the approximation of continuous random variables only at times when it is convenient;

- We consider negative exponents on phase separation and mixing operators, which by the symmetry of Grover mixer results exactly in the same expectation value as when assuming positive exponents, the definition used by Headley and Wilhelm [30];

- Headley and Wilhelm [30] assume on the analysis the context of unconstrained optimization. We, on the other hand, define GM-QAOA as acting in a generic feasible subspace. The equivalence is direct;

- As discussed previously, Headley and Wilhelm [30] applied the analysis to the Number Partition Problem. That example is not present in our work;

- Headley and Wilhelm [30] make a brief discussion about the circuit compilation of the Grover mixer operator, a topic not covered here;

- On the expectation value expression of arbitrary numbers of layers, we replace our original notation with the Headley and Wilhelm [30] because we find it more elegant. The same happens with the proof of Theorem 7;

- The original results of the present work are the following: the simplification of one layer case given Theorem 9 and Corollary 1; the exact number of terms of the general expectation value expression of GM-QAOA, given by Lemma 2; Corollary 3 and the discussion on standard score, which cover all Subsec. 4.1.3; the analytical results for the binary function on Subsec. 4.1.4; and the some of the numerical experiments and discussions for GM-QAOA of Chapter 5.

In his PhD thesis Ref. [33], David Headley includes some additional topics compared to the paper Ref. [30]: an alternative method to the Lemma 1 to get Eq. (4.7) from the last equality of Eq. (4.14); the conclusion that the performance of GM-QAOA is invariant over the two first statistical moments (we get the same conclusion here with a slightly different approach); the connection of GM-QAOA with Grover's search (which is deeper explored here on the binary function analysis of Subsec. 4.1.4); the use of the quantile of the expectation value as a metric of performance for GM-QAOA (here, we consider it from the analysis of the GM-Th-QAOA onwards); and numerical experiments with the continuous uniform distribution (numerically studied here), the triangle distribution, and unstructured search problem instances (i.e., bernoulli distributions)—in a comparative study that also includes the normal distribution, a probability distribution already considered in Ref. [30].

The second part of this work is the extension of the statistical approach to GM-Th-QAOA in Sec. 4.2 and the generalization of these results to a more general framework we called *Grover-based Quantum Alternating Operator Ansatz* or simply *Grover-based QAOA* in Chapter 5. These results are available on the preprint paper of Bridi and Marquezino [34]. Beyond the results on GM-Th-QAOA and Grover-based QAOA, we include in the paper some of the aforementioned original results of GM-QAOA, such as the application of the binary function and the discussion on the standard score.

The motivation for the analysis of GM-Th-QAOA is that although the statistical approach provides surprising simplifications in GM-QAOA expectation value calculations, the expression is still too complicated to obtain formal bounds on the algorithm's performance through direct analytical treatment. Thus, to understand the theoretical potential of the Grover mixer on combinatorial optimization—especially to investigate the issue concerning the quadratic speed-up over classical brute force—it is convenient to extend the analysis to the more simple case of GM-Th-QAOA. Using the well-known formula of probability of Grover's algorithm and its optimality on average probability for the unstructured search [35, 36], we provide an expression for the expectation value with complexity independent on the number of layers, which allows to study the asymptotic behavior of the algorithm. Rather than the characteristic function, the prominent statistical quantity here is the conditional expected value. With a closed-form expression, we prove the conjecture on which the efficient method of Golden et al. [26] of finding the optimization parameters is based (see Subsec. 3.8.2 and 4.2.1). Furthermore, we provide bounds on the performance of the expectation value of GM-Th-QAOA with the statistical quantities of quantile and the standard score. On the first, we get an asymptotic tight bound that implies the expected quadratic speed-up of GM-Th-QAOA over classical brute force. On the second, we conclude that the maximum standard score achieved by the expectation value of GM-Th-QAOA, hit by binary functions with specific ratios, scales linearly with the number of rounds. As an immediate consequence, we bound the number of layers to achieve a fixed approximation ratio. Finally, we combine both bounds to argue that the algorithm's performance is closely related to the asymptotic behavior of the probability distribution on the limit of its support.

To get stronger results about the Grover mixer, we introduced the more general Grover-based QAOA. In that framework, which includes both GM-QAOA and GM-Th-QAOA, is allowed that the phase separation operator codifies any compilation of the cost function in a real-valued function. We generalize all bounds of GM-Th-QAOA to Grover-based QAOA with a technique that consists of bounding the maximum amplification of the probability of measuring a set of degenerate states. The argument used for this is based on showing that if there is an amplification greater than the one provided by Grover's search, we contradict the optimality of average provability on unstructured search [35, 36], building an explicit algorithm that performance better than Grover's algorithm. All bounds on Grover-based QAOA follow the same asymptotic behavior as its correspondents on GM-Th-QAOA. As a consequence, we get the principal contribution of this work, the formalization of the notion that the Grover mixer is limited to the quadratic speed-up over classical brute force with an asymptotic performance, for instance, analog to the Grover Adaptive Search (GAS) [37–40]. We apply the bounds in the context of the Max-Cut problem. That way, by using the knowledge of the asymptotic behavior of the probability distribution associated with the particular case of complete bipartite graphs, we argue that for this class of graphs, the number of rounds required to achieve a fixed approximation ratio must grow exponentially with the number of vertices/edges, a severe limitation on the performance of the Grover mixer. More than that, the construction suggests that it is likely that the same happens with other classes of graphs and even with other combinatorial optimization problems.

1.2 Roadmap and general comments

The structure of the dissertation is as follows.

- In Chapter 2, we present a brief review of Probability Theory, highlighting important concepts for understanding this work;

- The Chapter 3 has the main objective of presenting GM-QAOA and GM-Th-QAOA. To get that, we present the more general context in which these variants are inserted— thereby reviewing the literature;

- In Chapter 4, we present the statistical approach, proving the analytical results of both GM-QAOA and GM-Th-QAOA;

- In Chapter 5, we provide numerical experiments involving probability distributions to emphasize and illustrate important aspects of the results of Chapter 4;
- In Chapter 6, we generalize the analytical bounds of GM-Th-QAOA to Grover-based QAOA and apply it to the Max-Cut problem;
- In Chapter 7, we present our conclusions and suggestions for future work.

Before proceeding, we need to make some general comments.

- We mention the work of Zhang et al. [41] in the conclusions of Chapter 7. That paper, released after the first version of Bridi and Marquezino [34] preprint, provides numerical evidence of quadratic speed-up of GM-QAOA over classical brute force;
- We emphasize a notation convention widely used throughout the present text from here. In particular, we distinguish between objects based on the presence of subscripts and superscripts; and by the present and the number of arguments. Some examples include: $\delta(x)$ denotes delta function, $\delta(x, y)$ denotes the Kronecker delta, and δ is a free symbol; the symbols $E_r(\boldsymbol{\beta}, \boldsymbol{\gamma})$, $E_r(t)$, and E_r denotes respectively the expectation value of GM-QAOA, GM-Th-QAOA, and Grover-based QAOA, while $\mathrm{E}[X]$ is the expected value of random variable X; $f_X(x)$ indicates the probability distribution of random variable X, $f_X^G(x)$ the generalized probability density function of a discrete random variable X, and $f(x)$ a free choice of function;
- All figures of this work were produced by the author. In particular, with the exception of Fig. 3.1 and Fig. 4.1(a), all graphics and numerical experimental (not limited to just Chapter 5) of this work were done using resources of *Python* [42], with the graphics specifically being produced with the *Python* package *Mathplotlib* [43];
- To get the numerical experiments of Chapter 5 (and other few situations), we need several statistical quantities, such as mean, standard deviation, and characteristic functions, for some families of probability distributions, such as the normal, Laplace, and gamma distributions. As the total number of used statistical quantities is large, it is unfeasible to demonstrate or cite all of them. Because of that, we assume that they are well-known. Most of them can be calculated using basic integration techniques, especially the quantity $G_X(x)$, defined on Subsec. 2.7.1, which is harder to find in the literature. An alternative for slightly more complicated cases is to compute it on algebraic software, such as *Maple* [44] or *Mathematica* [45].

Chapter 2

A Review on Probability Theory

The analysis method of QAOA with Grover mixer, addressed in Chapter 4, is based on Probability Theory. That way, in the present chapter, we do a brief review of this branch of mathematics, highlighting important concepts for understanding our results and introducing the notation to be used throughout this work. The main references used here are the book of Hoel, Port, and Stone [31] and Pishro-Nik's [32] book, using their definitions and concepts throughout this chapter. For specific topics, in addition to the papers and books cited throughout the text, the books of Spanos [46] and Sugiyama [47] are used especially but not limited to the topics of quantile and standardized moments, and Hoel, Port, and Stone [48] book for estimators and confidence intervals.

As the terminology suggests, Probability Theory deals with the mathematical concept of probability. Historically, it was developed to form the foundation for the mathematical treatment of random phenomena [49]. In contemporary times, it finds applications in numerous areas of knowledge, such as physics, biology, engineering, and finance [31]. Although there are different interpretations of the meaning of probability in the physical world, Probability Theory is independent of them, treating probability as a rigorous mathematical concept dependent only on its axioms [32]. In this chapter, we present Probability Theory from the point of view of mathematics, eventually citing some motivation from the physical world, such as rolling dice, to gain intuition. In particular, in Sec. 2.5, we present the concepts of estimators and confidence intervals, which are more associated with Statistics. Briefly, that field deals with the interpretation and inference of random phenomena from observed data, using techniques and mathematical models based on Probability Theory [48].

2.1 Probability spaces

The notion of probability as a numerical value between 0 and 1 to "something happens" is abstracted on probability theory by the concept of probability spaces.

In such mathematical objects, we assign *measures* with values on the aforementioned range to a set of abstract points termed sample space. That assignment is done through a real-valued function known as a probability measure. Formally, we define it as follows.

Definition 1 (Probability space) *A probability space is a tuple* $(\Omega, \mathcal{A}, \mathrm{P})$*, where* Ω *is the sample space;* $\mathcal{A}$ *is a nonempty collection*[1] *of subsets of* Ω *called events, that is a* σ*-field, i.e., that is closed under the set theory operations of complement, a countable number of unions, and countable number of intersections; and* P *is a probability measure defined on* $\mathcal{A}$*, denoted* $\mathrm{P}[A]$ *when applied to an event* A*, that satisfies the conditions*

- $\mathrm{P}[\Omega] = 1$;
- $\mathrm{P}[A] \geq 0 \ \forall A \in \mathcal{A}$;
- *If* A *is a countable collection of mutually disjoint events in* Ω*, then*

$$P\left[\bigcup_{a \in A} a\right] = \sum_{a \in A} \mathrm{P}[a]. \tag{2.1}$$

According to the discussion of Hoel, Port, and Stone [31], on the one hand, it is straightforward to define probability spaces for discrete sample spaces—i.e., on a countable subset of the real numbers—such as to model the experiment of drawing a ball from a box. On the other, the definition of probability space on continuous sample spaces (uncountable subset), as the modeling of the isotope disintegration experiment, is much deeper and involves questions answered by the advanced branch of measure theory. Fortunately, the results of measure theory, which are beyond the scope of this work, guarantee that constructions of continuous probability spaces are possible.

An example considered in that book is the important class of uniform probability spaces. In the discrete case, we can think of the intuitive notion of picking "at random" from a set C with c elements. That way, $\Omega = C$, $\mathcal{A}$ consist in all 2^c subsets of Ω, and $\mathrm{P}[A] = j/c$ for a event A with j elements. Each element has the same probability of $1/c$. The continuous case can be thought of as the experiment of choosing a point "at random" from an interval $C = [a, b]$ on the real line for $-\infty < a < b < \infty$. Now, we measure the "size" of an event C by its length, given by $b - a$ and denoted by $|C|$. Thus, by taking $\Omega = C$, measure theory results confirm the existence of a σ-field of subsets $\mathcal{A}$ of C, which in particular consist in all intervals of C. The probability measure defined on $\mathcal{A}$ is given by $\mathrm{P}[A] = |A|/|C|$ for any interval

[1] The terminology collection is used here as a synonym for set.

A. Of course, if A is a single point, $\mathrm{P}[A] = 0$ since $|A| = 0$. More generally, if C is a n-dimensional Euclidean space with finite and non-zero n-dimensional volume, it follows an analogous construction of probability space but with the measure of the n-dimensional volume instead of the particular case of length.

2.1.1 Properties of probability

We list some basic properties of the probability measure. Let A and B be two events and C a collection of events, then

- $\mathrm{P}[\varnothing] = 0$;
- $\mathrm{P}[A] = 1 - \mathrm{P}[\overline{A}]$[2] and
$$P\left[\bigcup_{c\in C} c\right] = 1 - P\left[\bigcup_{c\in C} \bar{c}\right]; \tag{2.2}$$
- $\mathrm{P}[B] = \mathrm{P}[A \cap B] + \mathrm{P}[\overline{A} \cap B]$;
- **(Monotonicity of probabilities)** $\mathrm{P}[B] \geq \mathrm{P}[A]$ if $A \subseteq B$;
- **(Inclusion-exclusion principle for 2 sets[3])** $\mathrm{P}[A \cup B] = \mathrm{P}[A] + \mathrm{P}[B] - \mathrm{P}[A \cap B]$;
- **(Union bound)**
$$P\left[\bigcup_{c\in C} c\right] \leq \sum_{c\in C} \mathrm{P}[c]. \tag{2.3}$$

2.1.2 Conditional probability

Consider the problem of finding the probability of an event given that another event occurred, such as the probability of an honest dice of 6 sides outcomes the number 5 given that the outcome is odd. That probability is called conditional probability, which is formally defined as follows.

Definition 2 (Conditional probability) *Let A and B be two events. Provided that $\mathrm{P}[A] > 0$, the conditional probability of B given A, is denoted by $\mathrm{P}[B|A]$ and given by*

$$\mathrm{P}[B|A] = \frac{\mathrm{P}[B \cap A]}{\mathrm{P}[A]}. \tag{2.4}$$

Intuitively, we can think of conditional expectation as a "cut" of the probability space. We consider the probability of both events occurring $\mathrm{P}[B \cap A]$ and then normalized by $\mathrm{P}[A]$ since we are restricted to the subset corresponding to the event A.

[2] $\overline{A}$ denotes the complement of A.

[3] There is a well-known formula for n sets.

2.1.3 Independence of events

Consider the case in which the knowledge that the event A such that $\mathrm{P}[A] > 0$ occurs does not affect the probability of an event B, i.e., $\mathrm{P}[B|A] = \mathrm{P}[B]$. By Eq. (2.4), $\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B]$, which can be trivially extended to the case of $\mathrm{P}[A] = 0$. By symmetry, we get the same result by interchanging the sets A and B. In this case, we say that the events are independent, which leads to the following definitions.

Definition 3 (Independence of 2 events) *Two events A and B are said to be independent if*

$$\mathrm{P}[A \cap B] = \mathrm{P}[A]\,\mathrm{P}[B]. \tag{2.5}$$

Definition 4 (Independence of n events) *The events $A_1, A_2, \ldots, A_n$ for $n \geq 3$ are said to be pairwise independent if every pair of events A_i and A_j with $1 \leq i < j \leq n$ are independent, and to be mutually independent if for any subcollection A of events containing at least two elements,*

$$P\left[\bigcap_{a \in A} a\right] = \prod_{a \in A} \mathrm{P}[a]. \tag{2.6}$$

Every collection of mutually independent events is pairwise independent, but a reciprocal is false.

Note that the formal definition of independence is different from its intuitive notion. For instance, one can verify that the events of taking a prime number and a number greater than 4 on an honest dice of 6 sides are independent.

2.2 Random variables

Instead of dealing with the explicit constructions of probability spaces, we can use the auxiliaries quantities called random variables, which are usually simpler and more convenient. Indeed, probability space are, in general, placed as background in the Probability Theory to make room for auxiliary quantities, as in the case of random variables [31].

Random variables are functions that assign values to outcomes of sample space. It can be discrete, continuous, or even mixed. As the terminology suggests, a mixed random variable contains discrete and continuous components. In this work, except for the definition of characteristic functions on Sec. 2.8, we restrict the random variables to real-valued functions. Although our following definition includes all aforementioned types of random variables, we consider only the discrete and continuous cases during this dissertation.

Definition 5 (Random variable) *A random variable X on a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ is a function $X(\omega) : \Omega \to \mathcal{R}$ in which $\{\omega : X(\omega) \leq x\}$ is an event such that for all $-\infty < x < \infty$.*

We simplify the notation of the event $\{\omega : X(\omega) \leq x\}$ as $\{X \leq x\}$. We distinguish between discrete and continuous random variables by the following definition.

Definition 6 (Discrete and continuous random variables) *A random variable X is discrete if the set of possible values of X is countable and continuous if*

$$\mathrm{P}[X = x] = 0, \ -\infty < x < \infty. \tag{2.7}$$

A mixed random variable, which combines discrete and continuous components, could be defined as a random variable with an uncountable set of possible values that do not satisfy Eq. (2.7). Immediate examples of discrete random variables are the constant random variable, defined by $X(\omega) = c$ for any ω where c is a real number, and the indicator random variable, which for an event A, $X(\omega) = 1$ if $\omega \in A$ and $X(\omega) = 0$ otherwise. To continuous case, we can cite the continuous uniform random variable, defined on $\Omega = [a, b]$ such that the event $\{X \leq x\}$ have probability $(x - b)/(b - a)$ if $x \in [a, b]$, 0 if $x < a$, and 1 if $x > b$.

2.2.1 Cumulative distribution function

A function closely related to our definition of random variables is the cumulative distribution function (cdf), defined as follows.

Definition 7 (Cumulative distribution function) *The cumulative distribution function of a random variable X, denoted $F_X(x)$[4], is given by*

$$F_X(x) = \mathrm{P}[X \leq x], \ -\infty < x < \infty. \tag{2.8}$$

The cdf presents the properties

- $0 \leq F_X(x) \leq 1$ for all x;
- $F_X(x)$ is a non-decreasing function of x;
- Taking the limits on $x \to -\infty$ and $x \to \infty$,

$$\lim_{x \to -\infty} F_X(x) = 0, \ \lim_{x \to \infty} F_X(x) = 1; \tag{2.9}$$

[4] In general, for the statistical quantities to be defined from here, we use the subscript to distinct between different random variables.

- For all x,

$$\lim_{k \to x^+} F_X(k) = F_X(x). \tag{2.10}$$

For any real-valued function $f(x)$ that satisfies these four properties, there is a probability space and a random variable X such that $F_X(x) = f(x)$ is the cdf.

An equivalent characterization for continuous random variables can be obtained in terms of the cdf as follows. Consider a related result to the fourth property given by

$$\lim_{k \to x^-} F_X(k) = \mathrm{P}[X < x] \tag{2.11}$$

for all x. Combining Eq. (2.10) and (2.11) gives, for all x,

$$\lim_{k \to x^+} F_X(k) - \lim_{k \to x^-} F_X(k) = \mathrm{P}[X \leq x] - \mathrm{P}[X < x] = \mathrm{P}[X = x], \tag{2.12}$$

which implies, by Def. 6, that X is continuous if and only if $F_X(x)$ is a continuous function.

Furthermore, the cdf can be used to compute the probability $\mathrm{P}[a < X \leq b]$ as

$$\mathrm{P}[a < X \leq b] = \mathrm{P}[X \leq b] - \mathrm{P}[X \leq a] = F_X(b) - F_X(a). \tag{2.13}$$

In particular,

$$\mathrm{P}[X > a] = 1 - F_X(a). \tag{2.14}$$

2.2.2 Probability distribution

In this work, we use the terminology probability distribution as a general term to combine the concepts of probability mass function (pmf) of discrete random variables and probability density function (pdf) of continuous random variables. Mass and density on pmf and pdf, respectively, are analogous to their use in physics [32]. In general terms, the probability distribution, denoted by $f_X(x)$, is a real-valued function whose output is related to the probability of the point x.

Probability mass function

For discrete random variables, the probability distribution is called probability mass function since we take directly the probability of the point x, which leads to the following definition.

Definition 8 (Probability mass function) *The probability mass function of a discrete random variable X is a real-valued function given by*

$$f_X(x) = \mathrm{P}[X = x], \ -\infty < x < \infty. \tag{2.15}$$

Given a pmf $f_X(x)$, we can calculate the cdf by the summation

$$F_X(x) = \sum_{k:f_X(k)>0, k\leq x} f_X(k). \tag{2.16}$$

Note that we could define discrete random variables based on their pmf, replacing on Def. 5 the event $\{\omega : X(\omega) \leq x\}$ by $\{\omega : X(\omega) = x\}$. However, that definition would not work for continuous random variables since these random variables must satisfy $\mathrm{P}[X = x] = 0$ for all x. A consequence of this condition is that the definition of probability distribution on continuous random variables needs to be different.

Probability density function

In the continuous case, instead of directly taking the value of probability as in the case of pmf, the probability distribution considers the probability per unit length, being because of it, known as the probability density function. Thus, taking the limit of length approaches 0 leads to

$$f_X(x) = \lim_{\Delta \to 0^+} \frac{\mathrm{P}[x < X \leq x + \Delta]}{\Delta} = \lim_{\Delta \to 0^+} \frac{F_X(x+\Delta) - F_X(x)}{\Delta} = \frac{F_X(x)}{dx}, \tag{2.17}$$

if the limit exists. Thus, we obtain the Def. 9.

Definition 9 (Probability density function) *The probability density function of a continuous random variable X is given by*

$$f_X(x) = \frac{dF_X(x)}{dx}, \ -\infty < x < \infty, \tag{2.18}$$

for all points in which $F_X(x)$ is differentiable.

Analogously to pmf, we can calculate the cdf from a $f_X(x)$ by the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(k)\ dk. \tag{2.19}$$

We also assume that the cdf of a continuous random variable is not differentiable at most in a finite set of points. In that case, we can arbitrate the respective values of $f_X(x)$ of these points without changing integral calculations.

General discussions

To condensate discrete and continuous random variables in a single case, we introduce the notation

$$\sum\!\!\!\!\!\!\int_{x \in A} f(x), \tag{2.20}$$

which means that for a subset A of the domain of the function $f(x)$, we sum the discrete values of x and integrate over the continuous interval concerning x. Similar notation is used, for instance, on the Deffner and Campbell [50] work.

To apply that notation in the context of random variables, we define $R_X = \{x : f_X(x) > 0\}$, a set known as the support of X. The minimum and the maximum values of the support R_X are denoted by R_X^{min} and R_X^{max}, respectively[5]. Thus, for instance, we can combine Eq. (2.16) and (2.19) to write $F_X(x)$ as

$$F_X(x) = \sumint_{k \in R_X : k \leq x} f_X(k), \tag{2.21}$$

or write the probability of an event A by

$$\mathrm{P}[X \in A] = \sumint_{x \in A} f_X(x). \tag{2.22}$$

The probability distribution has the properties $f_X(x) \geq 0$ for any $x \in \mathbb{R}$ and $\sumint_{x \in R_X} f_X(x) = 1$.

In this work, we name the random variable $-X$ by the reflected random variable of X. From the definition of cdf and from Eq. (2.14), follows that $F_{-X}(x) = 1 - F_X(-x) + \mathrm{P}[X = -x]$. Similarly, considering individually both pmf and pmf, we can conclude that $f_{-X}(x) = f_X(-x)$. That way, we can also use the terminology reflected probability distribution or simply reflected distribution. Moreover, if $f_X(x) = f_X(-x)$ for all x, we say that the random variable/distribution is symmetric.

Generalized probability distribution function

The concept of probability density function can be extended also to discrete (and mixed) random variables, which allows to unify of the theory of random variables [32]. To get that, note that since the cdf of a discrete random variable evolves through discrete jumps, it can be written by using the Heaviside step function $\theta(x)$, given by

$$\theta(x) = \begin{cases} 1, & x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.23}$$

Thus,

$$F_X(x) = \sum_{k \in R_X} f_X(k)\theta(x-k). \tag{2.24}$$

[5] If, for instance, $R_X = (-\infty, \infty)$ or $R_X = \mathbb{Z}$, then $R_X^{min} \to -\infty$ and $R_X^{max} \to \infty$.

The derivative of the step function is the delta function or delta Dirac function, a generalized function denoted by $\delta(x)$ and given by

$$\delta(x) = \begin{cases} \infty, & x = 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.25}$$

That way, the derivative of $F_X(x)$ is given by

$$\frac{dF_X(x)}{dx} = \sum_{k \in R_X} f_X(k)\delta(x - k), \tag{2.26}$$

which leads us to the next definition.

Definition 10 (Generalized probability density function) *The generalized probability density function for a discrete random variable X with probability mass function $f_X(x)$, denoted by $f_X^G(x)$, is given by*

$$f_X^G(x) = \sum_{k \in R_X} f_X(k)\delta(x - k). \tag{2.27}$$

The generalized pdf is particularly useful when we need to take the derivative of a discrete random variable, as in the case of the problem considered on Subsec. 4.2.1. However, in all other situations from here onwards, we consider the usual pmf to deal with discrete random variables.

2.2.3 Quantile function

The called quantile function, $Q_X(y)$, outputs a specific value of the random variable X such that the cdf evaluates in that value is greater or equal to y for $y \in [0, 1]$. For a continuous and strictly increasing cdf, the quantile is the inverse cdf, i.e., $Q_X(y) = F_X^{-1}(y)$. In the general case, to handle jumps and intervals of constant value on cdf, we extend the definition to

$$Q_X(y) = \min\{x \in \mathbb{R} : y \leq F_X(x)\}, \tag{2.28}$$

that is, we take the minimum value of x amongst all those values in which the cdf exceeds y. If $Q_X(y) = x$, we say that the value of x is at the yth quantile, or equivalently that $y = F_X(x)$ gives the quantile in which x is associated. Some important quantiles are the 0.5th, 0.25th, and 0.01th, known respectively as median, quartile, and percentiles. The quantile can be a useful metric for comparing distributions, as considered in Subsec. 4.2.2.

2.2.4 Random vectors

It is common in Probability Theory to be interested in examining the relationship of two or more random variables [31]. To deal with these issues, we can introduce the concept of random vectors. Let $X_1, \dots, X_n$ be a set of random variables. Then, the n-dimensional vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is called random vector. We also set the vector n-dimensional $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. The analog of cdf on random vectors, called joint cumulative distribution function (joint cdf), is given by

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{P}[X_1 \leq x_1, \dots, X_n \leq x_n]. \tag{2.29}$$

For probability distributions, the joint pmf and joint pdf are given analogously to the one-dimensional case by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \mathrm{P}[X_1 = x_1, \dots, X_n = x_n], \; f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{\partial^n F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \dots \partial x_n}, \tag{2.30}$$

respectively.

If we are interested in getting the individual probability distributions of random variables X, Y from a random vector (X, Y), called marginal probability distributions (marginal pmf/pdf), it can be showed analyzing individually the discrete and continuous cases that

$$f_X(x) = \sum\!\!\!\!\!\!\!\int_{y \in R_Y} f_{X,Y}(x, y), \; f_Y(y) = \sum\!\!\!\!\!\!\!\int_{x \in R_X} f_{X,Y}(x, y). \tag{2.31}$$

In an analogous way, for a n-dimensional random vector, the marginal probability distribution of a particular random variable X_k such that $1 \leq k \leq n$, is

$$f_{X_k}(x_k) = \sum\!\!\!\!\!\!\!\int_{x_1 \in R_{X_1}} \cdots \sum\!\!\!\!\!\!\!\int_{x_{k-1} \in R_{X_{k-1}}} \sum\!\!\!\!\!\!\!\int_{x_{k+1} \in R_{X_{k+1}}} \cdots \sum\!\!\!\!\!\!\!\int_{x_n \in R_{X_n}} f_{\boldsymbol{X}}(\boldsymbol{x}). \tag{2.32}$$

The marginal cdf can be obtained by simply taking the limit of the remainder variables to approach infinity, i.e.,

$$F_{X_k}(x_k) = \lim_{x_1 \to \infty} \dots \lim_{x_{k-1} \to \infty} \lim_{x_{k+1} \to \infty} \dots \lim_{x_n \to \infty} F_{\boldsymbol{X}}(\boldsymbol{x}). \tag{2.33}$$

2.2.5 Independence of random variables

Consider the experiment of rolling a dice two times. Intuitively, we have the notion that the output of the first dice does not affect the result of the second one [31]. This notion can be precisely stated by saying that two random variables X and Y

are called independent if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y). \tag{2.34}$$

for all x, y. This definition is generalized to n random variables in Def. 11.

Definition 11 (Independence of random variables) *The random variables X_1, $X_2, \ldots, X_n$ are said to be mutually independent, or simply independent, if*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1}(x_1)\ldots F_{X_n}(x_n), \tag{2.35}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$.

If a set of independent random variables share the same cdf, we say that they are independent and identically distributed (i.i.d.) random variables.

2.2.6 Sum of independent random variables

Many algebraic operators can be performed with random variables, from the sum or multiplication of a random variable by a scalar, a topic covered in Sec. 2.4, to arithmetic operations between random variables. Now, in particular, we consider the sum of independent random variables. By considering individually both discrete and continuous, we can conclude for independent random variables X and Y that

$$f_{X+Y}(x) = \sum\!\!\!\!\!\!\int_{y\in R_X} f_X(y)f_Y(x-y), \ -\infty < x < \infty. \tag{2.36}$$

That operation is a convolution, denoted by $[f_X * f_Y](x)$. In general, for independent random variables $X_1, \ldots, X_n$ holds

$$f_{X_1+\ldots+X_n}(x) = [f_{X_1} * \ldots * f_{X_n}](x), \ -\infty < x < \infty. \tag{2.37}$$

2.3 Families of probability distributions

In this section, we present all probability distributions (continuous and discrete) relevant to this work, encompassing some of the most important probability distributions of the literature. although we abbreviate by probability distributions, they are families of parameterized probability distributions. In general, we introduce the distributions directly together with its support instead of defining explicitly random variables.

- **(Degenerate distribution)** We begin with the trivial degenerate distribution, denoted by Degenerate(c)[6], for $c \in \mathbb{R}$, with pmf given by $f_X(c) = 1$[7]. Note that Degenerate(c) is the distribution of the constant random variable. We also refer to degenerate distribution as a single-point distribution.

- **(Bernoulli distribution)** The distribution Bernoulli(p) with $0 \le p \le 1$ has pmf given by

$$f_X(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0. \end{cases} \tag{2.38}$$

  That distribution can be seen intuitively as the distribution of tossing a coin with arbitrary probability p. Note that the random variable associated with the Bernoulli distribution is the indicator random variable and that $p = 0$ or $p = 1$ reduces it to a degenerate distribution. The Bernoulli distribution is a particular case of the family of two-point distributions, defined similarly to the Bernoulli distribution but in which the support is given by arbitrary points a and b instead of 0 and 1.

- **(Binomial distribution)** The distribution Binomial(n,p) with $0 \le p \le 1$ and $n \in \{0,1,2,\dots\}$ is the sum of n i.i.d. random variables with distribution Bernoulli(p). The pmf is given by

$$f_X(x) = \binom{n}{x} p^k (1-p)^{n-x}, \; x \in \{0,1,\dots,n\}. \tag{2.39}$$

- **(Discrete uniform distribution)** The probability distribution DUniform(a,b) with $a,b \in \mathbb{Z}$ such that $b \ge a$ and pmf

$$f_X(x) = \frac{1}{b-a+1}, \; x \in \{a, a+1, \dots, b-1, b\} \tag{2.40}$$

  is a discrete example of uniform probability space.

- **(Continuous uniform distribution)** The distribution CUniform(a,b) with $-\infty < a < b < \infty$ has pdf given by

$$f_X(x) = \frac{1}{b-a}, \; x \in [a,b]. \tag{2.41}$$

  That distribution is a continuous example of uniform probability space.

[6] We follow that pattern of notation for probability distributions. To denote the reflected distribution, we distinguish with the letter R. For instance, RDegenerate(c) is the reflected version of the degenerate distribution.

[7] To simplify notation, we ignore the values of x such that $x \notin R_X$ in the writing of pmf.

- **(Normal distribution)** The normal distribution is a distribution of unique importance, for reasons that should be clear when we discuss the central limit theorem in Subsec. 2.5.2. Denoted by $\text{Normal}(u, s^2)$ with $u \in \mathbb{R}$ and $s > 0$, its pdf is given by

$$f_X(x) = \frac{1}{s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{s}\right)^2},\ x \in (-\infty, \infty). \tag{2.42}$$

Eq. (2.42) is also well-known as the Gaussian function.

- **(Discrete Gaussian kernel (DGK))** To discretize the normal distribution, one can simply consider a discrete support for the Gaussian function. A more refined approach is the discrete Gaussian kernel [51, 52]. While the Gaussian function is a solution of the diffusion equation with continuous time and space, DGK is the correspondent solution for the diffusion equation with discrete space (and continuous time). The distribution $\text{DGK}(s^2)$ has pmf is given by

$$f_X(x) = e^{s^2} I_x(s^2),\ x \in \mathbb{Z}, \tag{2.43}$$

where $I_x(s^2)$ is the modified Bessel function of integer order x [53].

- **(Chi-squared distribution)** The chi-squared distribution raises naturally when we consider the squared of the normal distribution. More precisely, if $X_1, \ldots, X_k$ are i.i.d. random variables with distribution $\text{Normal}(0,1)$ and $k \in \{1, 2, \ldots\}$, then the distribution of the sum $X_1^2 + \ldots + X_k^2$ is the $\text{Chi-squared}(k)$ of k degree of freedom, which has pdf

$$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \tag{2.44}$$

with $R_X = (0, \infty)$ if $k = 1$ and $R_X = [0, \infty)$ otherwise. The function $\Gamma(z) = \int_0^\infty \tau^{z-1} e^{-\tau}\, d\tau$ is the well-known gamma function, the extension of factorial to complex numbers.

- **(Gamma distribution)** $\text{Gamma}(a, b)$ with $a, b > 0$ is a generalization of the chi-squared distribution. In particular, $\text{Chi-squared}(k)$ is given by setting $a = k/2$ and $b = 1/2$. The pdf of the gamma distribution is given by

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \tag{2.45}$$

with $R_X = (0, \infty)$ if $k \leq 1$ and $R_X = [0, \infty)$ otherwise.

- **(Exponential distribution)** The distribution $\text{Exponential}(l)$ with $l > 0$ is another special case of gamma distribution. In this case, we set $a = 1$ and

$b = l$, which gives the pdf

$$f_X(x) = le^{-lx},\ \ x \in [0,\infty). \tag{2.46}$$

- **(Laplace distribution)** The distribution Laplace(u, b) with $u \in \mathbb{R}$ and $b > 0$ has pdf given by
$$f_X(x) = \frac{1}{2b}e^{-\frac{|x-u|}{b}},\ \ x \in (-\infty,\infty). \tag{2.47}$$
The difference between two i.i.d. Exponential(l) random variables has a distribution Laplace$(0, 1/l)$.

- **(Logistic distribution)** The distribution Logistic(u, s) with $u \in \mathbb{R}$ and $s > 0$ has pdf given by
$$f_X(x) = \frac{e^{-(x-u)/s}}{s(1+e^{-(x-u)/s})^2},\ x \in (-\infty,\infty). \tag{2.48}$$

- **(Pareto distribution)** The distribution Pareto(α, x_m) with $\alpha, x_m > 0$ has pdf given by
$$f_X(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}},\ \ x \in [x_m,\infty). \tag{2.49}$$
Note that the decay of that distribution is rigid by a power-law relationship, a characteristic that is important for the discussion of Subsec. 4.2.4.

2.4 Moments

The moments, or statistical moments, are numerical values used to make educated guesses regarding the form of probability distributions [46]. To introduce the moments, we begin discussing the two main moments, the expectation and the variance, to later generalize the concept. As noticed by Hoel, Port, and Stone [31], the general definition of expectation has details that require further background in the theory of measure and integration, which, in particular, is beside the point of this work. That way, is enough to present the "computational" Def. 12. Furthermore, since the moments are defined in terms of the expectation, the conclusion is the same in general.

2.4.1 Expectation

The expectation[8] of a random variable, denoted $\mathrm{E}[X]$, is a sum/integral of its possible values weighted by their respective mass/density of probability, i.e., $\mathrm{E}[X] =$

[8] Also named expected value, average, or mean.

$\SumInt_{x \in R_X} x f_X(x)$. That definition is valid if the sum/integral is well-defined, that is, $\SumInt_{x \in R_X} |x| f_X(x) < \infty$.

Definition 12 (Expectation) *Let X be a random variable. Provided that $\SumInt_{x \in R_X} |x| f_X(x) < \infty$, we say that X have finite expectation, defined by*

$$\mathrm{E}[X] = \SumInt_{x \in R_X} x f_X(x). \tag{2.50}$$

Other notation for the mean is μ_X. Some important basic properties of the expected value are listed.

- Expectation is linear;
- $|\mathrm{E}[X]| \leq \mathrm{E}[|X|]$;
- If for some constant c, $\mathrm{P}[|X| \leq c] = 1$, then X has finite expectation and follows the inequality $|\mathrm{E}[X]| \leq c$;
- For independent random variables X and Y with finite expectation, $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$.

A crucial result involving expected value is the law of the unconscious statistician (LOTUS), which calculates the expectation of a real-valued function $f(X)$.

Theorem 1 (LOTUS) *Let $f(x)$ be a real-valued function defined on the real line. If $\SumInt_{x \in R_X} |f(x)| f_X(x) < \infty$, then*

$$\mathrm{E}[f(X)] = \SumInt_{x \in R_X} f(x) f_X(x). \tag{2.51}$$

2.4.2 Variance and standard deviation

As discussed by Hoel, Port, and Stone [31], the expected value is tentative to summarize a probability distribution by a number representing its "typical value". The quality of that information depends on how clustered the values are about the expected value. A quantity that measures that spread from the mean is the variance, defined as follows.

Definition 13 (Variance and standard deviation) *Let X be a random variable. If X^2 has a finite expectation, then*

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mu_X)^2] \tag{2.52}$$

is a non-negative number called the variance of X, and we say that X has finite variance. The number $\sigma_X = \sqrt{\mathrm{Var}[X]}$ is the standard deviation of X.

The variance can be computed using LOTUS. Furthermore, from the linearity of expectation follows that $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. Other basic facts about variance are the following.

- $\mathrm{Var}[X] = 0$ if and only if X is a constant random variable;
- For $a, b \in \mathbb{R}$, $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$, $\mathrm{Var}[X + b] = \mathrm{Var}[X]$, and $\mathrm{Var}[X] = \mathrm{Var}[-X]$;
- For independent random variables $X_1, \dots, X_n$ with finite variance such that $X = X_1 + \dots + X_n$, $\mathrm{Var}[X] = \sum_{j=1}^{n} \mathrm{Var}[X_j]$.

2.4.3 Generalizing the moments

The expectation and the variance are known as the first and the second moments, respectively. We can extend it and define the nth moment. Indeed, there are (at least) three distinct types of moments: the raw moment, the central moment, and the standardized moment.

Definition 14 (Moments) *Let X be a random variable. If X^n has a finite expectation, we say that X has the nth moment or the moment of order n. In this case, the nth raw moment, the nth central moment, and the nth standardized moment are defined by* $\mathrm{E}[X^n]$, $\mathrm{E}[(X - \mu_X)^n]$, *and* $E\left[\left(\frac{X-\mu_X}{\sigma}\right)^n\right]$, *respectively.*

We can compute the moments via LOTUS. If X has the nth moment, then it also has the kth moment for all $k \leq n$. Moreover, if the random variables X and Y have the nth moment, then $X + Y$ also does have.

The mean is the first raw moment, or simply the first moment since both first central and standardized moments trivially are 0. The variance is the second central moment. Here, we refer to variance as just the second moment because it has greater theoretical and practical importance than the second raw moment $\mathrm{E}[X^2]$. Furthermore, the second standardized moment is trivially 1.

Location shifting and changing the scale

The first and second moments are related to the location and the scale, respectively, of the probability distribution. To see it, let X be a random variable and $a \in \mathbb{R}$. Directly from the definition of cdf, pmf, and pdf, $F_{X+a}(x) = F_X(x - a)$ and $f_{X+a}(x) = f_X(x - a)$, that is, a location shifting on probability distribution. As $\mathrm{E}[X + a] = \mathrm{E}[X] + a$ and $\sigma_{X+a} = \sigma_X$, the location shifting affects the first but not the second moment. Now, consider $b > 0$. Then, $F_{bX}(x) = F_X(x/b)$, $f_{bX}(x) = f_X(x/b)$ for pmf, and a simple substitution on derivative shows that the pdf is given by $f_{bX}(x) =$

$\frac{1}{b}f_X(x/b)$, which means a change of the scale. Intuitively, the factor $1/b$ on pdf can be seen as a normalization for the area under the curve on probability distribution to keep it equal to 1, which, of course, is not necessary on pmf. As $\mathrm{E}[bX] = b\,\mathrm{E}[X]$ and $\sigma_{bX} = b\sigma_X$, it affect both first and second moments if $\mathrm{E}[X] \neq 0$. Usually, families of distributions have parameters that correspond to location shifting and changing the scale. For instance, in distribution Normal(u, s^2), u and s^2 concerning location and scale, respectively, being, in that particular case, the mean and the variance of the distribution.

To combining both cases, let Y and Z be random variables such that $X = Y + \mu_X$ and $X = \sigma_X Z + \mu_X$. Follows that $\mathrm{E}[Y] = \mathrm{E}[Z] = 0$ and $\sigma_Z = 1$. The random Y changes the location to make the mean equal to 0. Because of that, we can extend the definition of symmetric random variable/distribution to say that if $f_Y(x) = f_Y(-x)$ for all x, the random variable/distribution is symmetric around the mean. The random variable Z additionally changes the scale to get unit standard deviation, therefore being invariant under the two first moments. For that reason, we call it in this work by the standard random variable associated with X. For named distributions, we call, for instance, the standard version of normal distribution as standard normal distribution. A value z of the support R_Z is called standard score [54]. The correspondence of z with an element x of R_X is given by $z = (x - \mu_X)/\sigma_X$, which means it measures the number of standard deviations away from the mean. As well as quantile, the standard score, can be used as a metric to compare different distributions, as done in Subsections 4.1.3 and 4.2.3, and in Chapter 6. Note that the three defined types of moments are directly related to X, Y, and Z. The nth raw, central, and standardized moments are given by $\mathrm{E}[X^n]$, $\mathrm{E}[Y^n]$, and $\mathrm{E}[Z^n]$, respectively. Fig. 2.1 illustrates the location shifting and the change of the scale on random variables X, Y, and Z for the normal distribution.

Due to its importance for this work, we address the effect of location shifting and changing the scale for statistical quantities of interest defined from now on. The same holds with the reflected distributions.

Changing the shape

From the third moment onwards, it is reasonable to give the protagonism on the discussion of statistical moments to the standardized moments. The reason behind this is that, as discussed, the first and the second moments are related, respectively, to the location and scale of the probability distribution. Moments of higher order, on the other hand, concern the shape of the distribution. In that way, the standardized moment is usually the ideal choice, once it abstracts the location and scale by taking the corresponding random variable with zero mean and unit standard score. Some families of distributions have a parameter of shape, as gamma distri-

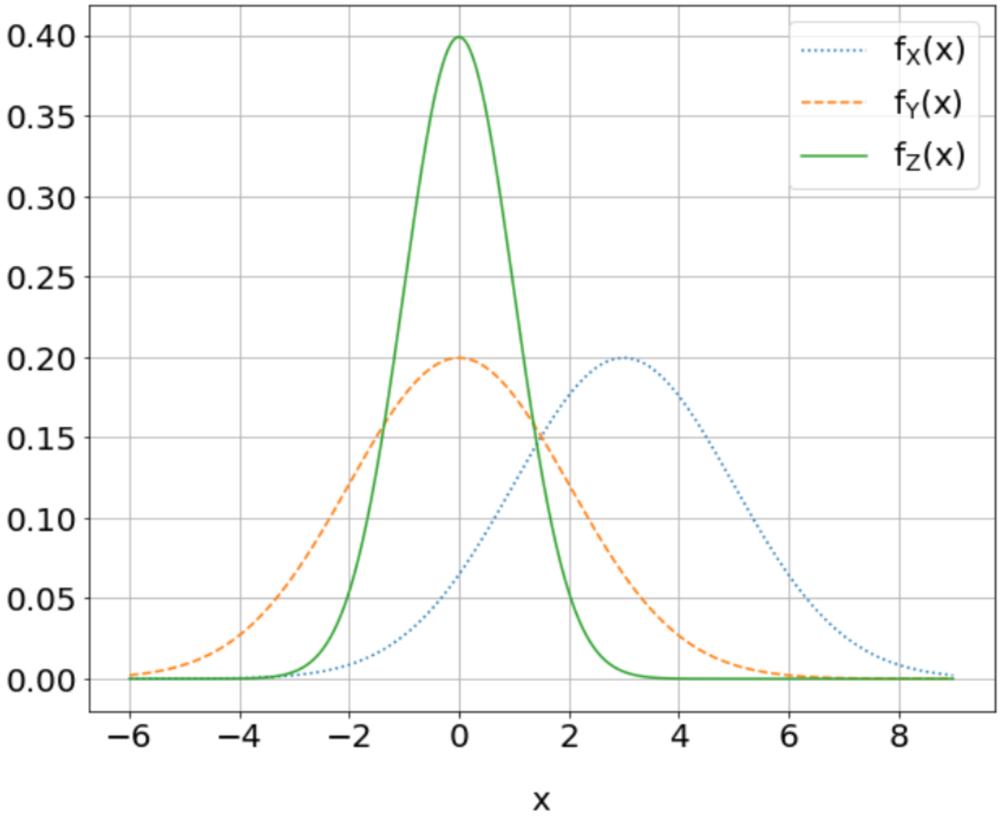


Figure 2.1: Probability density functions of the random variable X given by the distribution Normal(u, s^2) with $u = 3$ and $s = 2$, and its associated Y and Z random variables. The random variable Y is a location shifting to get zero expectation, while Z is additionally a change of the scale to get a unit standard deviation.

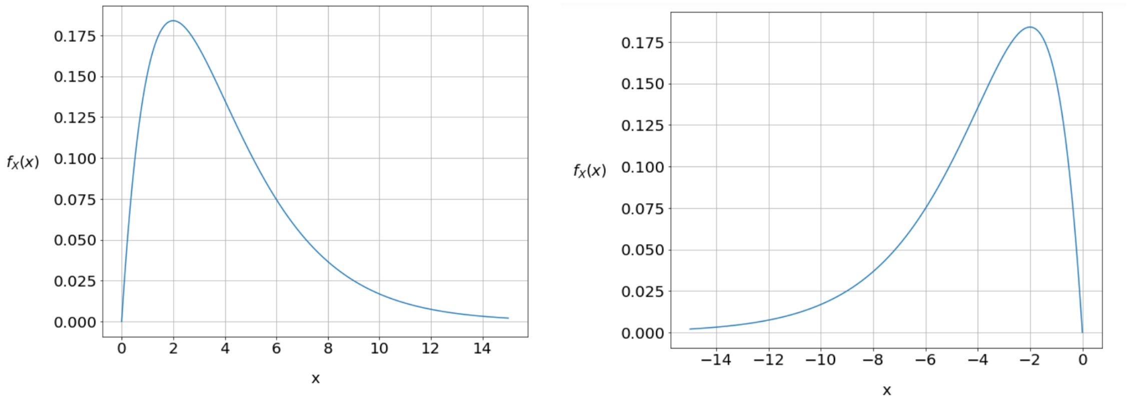
bution with parameter a. Such distributions have variable standardized moments, while distributions without shape parameters, as the normal distribution, have fixed standardized moments.

The third moment measures the asymmetry of the probability distribution about its mean, with the standardized moment known as skewness and denoted as Skew$[X]$. Since we exponentiate Z with an odd power, the points below the mean contribute negatively to the sum/integral, while points above the mean contribute positively in such a way that the skewness, if exists, can be positive, negative, or zero. That naturally induces a graphic interpretation of the skewness by the tails of distribution, i.e., the appendages on the sides of a distribution. If the right tail is longer than the left tail, the mass/density is concentrated on the right side of the distribution, and the skewness tends to be positive. On the other hand, if the left tail is longer, the concentration is on the left side, and the skewness tends to be negative. Because of this, we say that the distribution is right-skewed or skewed to the right if has positive skewness and left-skewed or skewed to the left for negative skewness. Furthermore, if a probability distribution $f_X(x)$ is right-skewed, than the reflected distribution $f_{-X}(x)$ is left-skewed since $\text{Skew}[X] = \text{Skew}[-X]$, and vice-versa.

Two aspects of the skewness must be considered. First, a distribution that is symmetric around its mean has $\text{Skew}[X] = 0$, but the converse, in general, is not true, since skewness is not linear. For instance, one can verify that the non-symmetric distribution

$$f_X(x) = \begin{cases} 0.4, & x = -2 \\ 0.5, & x = 1 \\ 0.1, & x = 3, \end{cases} \tag{2.53}$$

has $\text{Skew}[X] = 0$ [46]. Second, in right-skewed distribution, the mean is often larger than the median, while in left-skewed distributions, the mean is often smaller than the median. However, as well as the issue of $\text{Skew}[X] = 0$ and symmetric distributions, that is not always valid, as discussed in von Hippel's [55] paper. On the other hand, a distribution symmetric around its mean always has a median equal to the mean. Both observations emphasize the limitation of the interpretation of skewness and the importance of looking not only at the moments but also at the graphic of the distribution—also emphasized by Spanos [46]. Furthermore, in the particular case of the third moment, it raises the possibility of alternative metrics for measuring the asymmetry, such as the Pearson formula [55], that considers the difference between mean and median normalized by the standard deviation. Fig. 2.2 illustrates the graphic interpretation of positive and negative skewness for a distribution in which it holds.



(a)

(b)

Figure 2.2: Graphic interpretation of skewness. (a) The right-skewed distribution Chi-squared(k) with 4 degrees of freedom. The skewness is $\sqrt{2}$, the mean 4, and the median ≈ 3.5679. (b) The left-skewed RChi-squared(k) with 4 degrees of freedom. The skewness is $-\sqrt{2}$, the mean -4, and the median ≈ -3.5679. These specific distributions follow the described "rule" that relates skewness and mean/median positions.

The fourth moment, with a standardized moment named by kurtosis and denoted Kurt$[X]$, should be interpreted more carefully. Historically, kurtosis has been incorrectly interpreted as a measure of "peakedness" of the distribution, i.e., high kurtosis indicates a distribution with a higher peak, and low kurtosis is a flatter distribution [56]. In 2014, Westfall [56] settled the issue, establishing that the interpretation of kurtosis must be unequivocally in terms of tail extremity, i.e., the propensity to produce outliers. From the definition of kurtosis, we exponentiate Z at the fourth power. Thus, differently from skewness, all values contribute positively from the sum/integral, and consequently, outliers have a stronger impact. Fig. 2.3 illustrates the impact of outliers on kurtosis comparing distinct distributions.

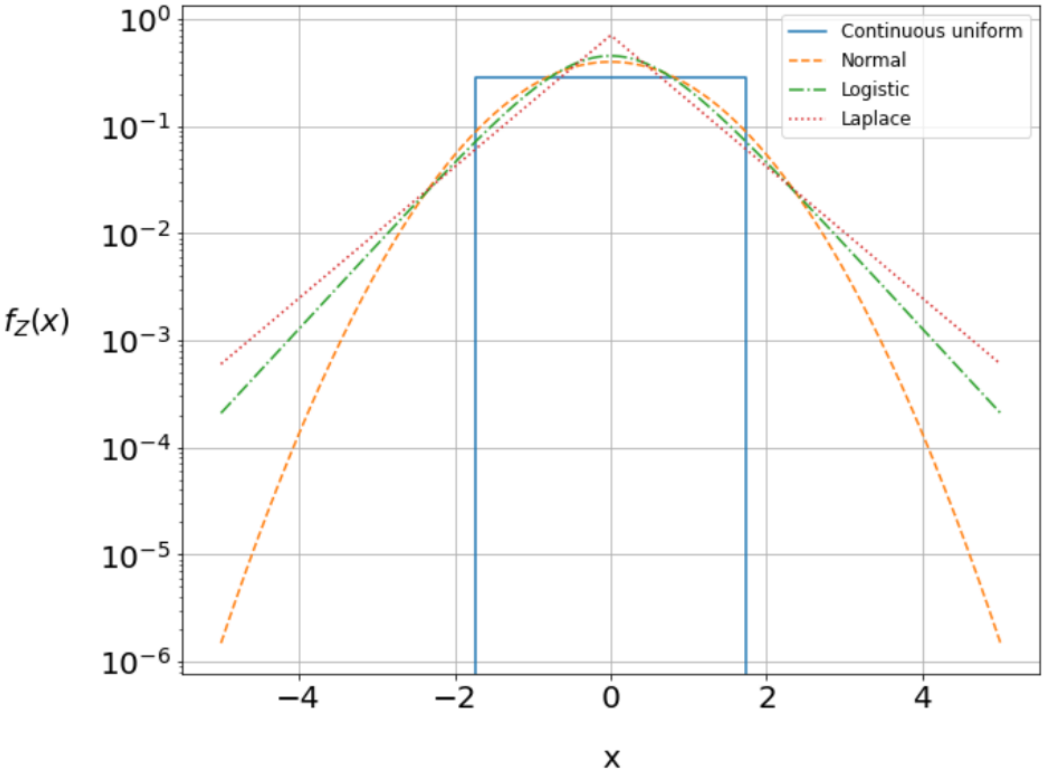


Figure 2.3: Log-linear graphic of the pdf of the standardized distributions $f_Z(x)$ of continuous uniform, normal, logistic, and Laplace distributions. All of them are symmetric around the mean distributions without a shape parameter. The choice of parameters to get Z gives CUniform$(-\sqrt{3}, \sqrt{3})$, Normal$(0, 1)$, Logistic$(0, \sqrt{3}/\pi)$, and Laplace$(0, \sqrt{2}/2)$, with the kurtosis given respectively by $9/5$, 3, $21/5$, and 6. The continuous uniform distribution, since is zero for $|x| > \sqrt{3}$, has a smaller impact on outliers and presents the lowest kurtosis between the distributions. To the remainder distributions, from a certain value of $|x|$, the pdf of normal distribution decays more quickly than logistic distribution, which in turn decays faster than Laplace distribution. Therefore, the increasing order of the impact of outliers, given by normal, logistic, and Laplace distributions, fits with the increasing order of their kurtosis values.

The minimum kurtosis is 1, hit by any two-point distribution with ratio $p = 0.5$—in particular, the Z random variable of two-point distribution depends only on the ratio (see Subsec. 2.6.2 for a proof). The normal distribution, which has kurtosis equal to 3, was historically used as a reference point of measure. Subtracting 3 from kurtosis gives the called excess kurtosis. Distributions with zero excess kurtosis are called mesokurtic, while distributions with positive and negative excess kurtosis are called leptokurtic and platykurtic, respectively. Furthermore, since kurtosis is an even moment, $\mathrm{Kurt}[X] = \mathrm{Kurt}[-X]$.

Naturally, one can generalize the arguments of the above discussion to moments of order above 4. Odd moments above skewness also measure the asymmetry but with an increasingly greater impact on the outliers. Even moments above kurtosis also measure the tail extremity with an increasingly greater impact on the outliers.

2.5 Random sample

We say that n i.i.d. random variables, each with the cdf $F_X(x)$, form a random sample[9] of size n from a *population* characterized by $F_X(x)$. We call here the process of taking a random sample from a population by random sampling. We also refer to sampling over a uniform sample space as uniformly sampling or choosing uniformly at random.

Let $X_1, \dots, X_n$ be a random sample. If we order the random variables in ascending order and denote the resulting sequence of random variables by $X^{(1,n)}, \dots, X^{(n,n)}$, follows that $X^{(1,n)} = \min\{X_1, \dots, X_n\}$ and $X^{(n,n)} = \max\{X_1, \dots, X_n\}$, and we say that $X^{(k,n)}$ is the kth order statistics of the random sample of size n.

2.5.1 Estimators and confidence intervals

In Statistic, an estimator is a function $d(X_1, \dots, X_n)$ of a random sample $X_1, \dots, X_n$ for estimate an unknown parameter $\hat{\theta}$ of the population. We call the numerical value resulting from observable values by an estimate of $\hat{\theta}$. An estimator for a parameter $\hat{\theta}$ is said to be unbiased if $\mathrm{E}[d(X_1, \dots, X_n)] = \hat{\theta}$ for any value of $\hat{\theta}$.

Estimators of particular interest are the sample moments, which are estimators for the population moments. For a random sample $X_1, \dots, X_n$, the kth sample raw, central and standardized moments are given by

$$m_k^r = \frac{1}{n}\sum_{j=1}^{n} X_j^k,\ m_k^c = \frac{1}{n}\sum_{j=1}^{n}(X_j - \hat{\mu}_X)^k,\ m_k^s = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{X_j - \hat{\mu}_X}{\hat{\sigma}_X}\right)^k, \tag{2.54}$$

respectively, where $\hat{\mu}_X = m_1^r$ is the sample mean and $\hat{\sigma}_X^2 = m_2^c$ is the sample

[9] With replacement, of course.

variance—$\hat{\sigma}_X$ is the sample standard deviation. The sample mean is unbiased, while the sample variance is not[10].

A way to estimate the accuracy of estimators is by building a confidence interval, which gives a bound on the probability of an estimator being on a specific interval. We say that a confidence interval has a confidence level $1 - \alpha$ on an interval $[\hat{\theta}_{min}, \hat{\theta}_{max}]$ if

$$\mathrm{P}[\hat{\theta}_{min} \leq \hat{\theta} \leq \hat{\theta}_{max}] \leq 1 - \alpha. \tag{2.55}$$

We consider here as an example the confidence interval of the sample mean for a population with distribution Normal(u, s^2). The pdf of the sample mean is the distribution Normal$(u, s/\sqrt{n})$. Thus, we have the confidence interval of

$$P\left[\hat{\mu}_X - z\frac{\hat{\sigma}_X}{\sqrt{n}} \leq \hat{\theta} \leq \hat{\mu}_X - z\frac{\hat{\sigma}_X}{\sqrt{n}}\right] = 1 - \alpha, \tag{2.56}$$

where z is the standard score associated with the confidence level, which can be obtained with tables or numerical methods. For instance, $z = 1$ is associated with $\alpha \approx 0.3173$ and $z = 2$ with $\alpha \approx 0.0455$. As discussed in detail by Hoel, Port, and Stone [48], a confidence interval for the sample variance on a population normally distributed can be built in terms of the chi-squared distribution.

2.5.2 Central limit theorem

A result related to random samples is the central limit theorem (CLT), which is one of the most remarkable theorems in Probability Theory [31]. It states that the sum of n i.i.d. random variables with finite second moment approaches the normal distribution when $n \to \infty$.

Theorem 2 (Central limit theorem) *Let $X_1, \ldots, X_n$ be i.i.d. random variables with non-zero finite variance. Set $S_n = X_1 + \ldots + X_n$ and Z_n as the standard random variable of S_n. Thus,*

$$\lim_{n\to\infty} F_{Z_n}(x) = F_N(x), \ -\infty < x < \infty, \tag{2.57}$$

where N is a random variable with standard normal distribution.

The CLT explains, at least partially, the ubiquitousness of the normal distribution in the fields of Probability Theory and Statistics. The sum of i.i.d. random variables appears in many situations, such as when an experiment is repeated assuming ideal conditions. Note that, for instance, the distributions Binomial$(n,\ p)$

[10] The sample variance would be unbiased if we defined it as $\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \hat{\mu}_X)^2$.

and Chi-squared(k) approaches normal distribution when $n \to \infty$ and $k \to \infty$, respectively. Naturally, as noticed by Hoel, Port, and Stone [31], the CLT suggests the approximation $F_{Z_n}(x) \approx F_N(x)$ for large n to compute the sum of arbitrary i.i.d. random variables, a useful formula in practical applications.

2.6 Probability bounds

Many probability inequalities are useful not just for proving theorems on Probability Theory but also for applied Statistics, especially when the exact calculation is complicated or the probability distribution is unknown [32]. An example of probability bound is the union bound, shown on Subsec. 2.1.1. In this section, we first show the Markov and Chebyshev inequalities, and then the Jensen's inequality. However, we can cite others, such as the Chernoff bounds, and the Cauchy-Schwarz and Hölder inequalities.

2.6.1 Markov and Chebyshev inequalities

Markov and Chebyshev inequalities bound the probability in terms of the first and second moments, respectively. For the first one, let X be a non-negative random variable. Then, for a given $a > 0$,

$$\mathrm{E}[X] = \sum\!\!\!\!\!\!\!\int_{x \in R_X} x f_X(x) \geq \sum\!\!\!\!\!\!\!\int_{x \in R_X : x \geq a} x f_X(x) \geq \sum\!\!\!\!\!\!\!\int_{x \in R_X : x \geq a} a f_X(x) = a\,\mathrm{P}[X \geq a], \quad (2.58)$$

and we get the following result.

Theorem 3 (Markov's inequality) *If X is a non-negative random variable and $a > 0$, then*

$$\mathrm{P}[X \geq a] \leq \frac{\mathrm{E}[X]}{a}. \quad (2.59)$$

Applying Marvok's inequality to a random variable $(X - \mu_X)^2$ and a constant b^2 proves Chebyshev's inequality, given by the following theorem.

Theorem 4 (Chebyshev's inequality) *If X is random variable and $b > 0$, then*

$$\mathrm{P}[|X - \mu_X| \geq b] \leq \frac{\mathrm{Var}[X]}{b^2}. \quad (2.60)$$

Chebyshev's inequality tends to provide better bounds than Markov's inequality, while the aforementioned Chernoff bounds tend to overcome both.

2.6.2 Jensen's inequality

The Jensen's inequality uses the concept of convex function. As noticed by Pishro-Nik's [32], intuitively, a function is considered convex when, upon selecting two points from its graph and connecting them with a line segment, the entire segment is positioned above the graph. Formally, we define it as follows. Let a function $f(x) : A \to \mathbb{R}$ with A being an interval on the real line. If, for any pair of points x and y on the interval A and any ρ between 0 and 1,

$$f(\rho x + (1 - \rho)y) \leq \rho f(x) + (1 - \rho)f(y), \tag{2.61}$$

we say that the function $f(x)$ is convex. So, Jensen's inequality states the following.

Theorem 5 (Jensen's inequality) *If $f(x)$ is a convex function on the support R_X of a random variable X in which $\mathrm{E}[f(x)]$ and $f(\mathrm{E}[X])$ are finite, then $\mathrm{E}[f(x)] \geq f(\mathrm{E}[X])$. The equality is hit if and only $f(x)$ is affine or X is a constant random variable.*

An immediate application of Jensen's inequality is to prove the trivial fact that $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 \geq 0$, which follows since x^2 is a convex function. Note that we also prove the property that $\mathrm{Var}[X] = 0$ if and only if X is a constant random variable. Another application is to prove that the minimum value of kurtosis, hit by two-point distributions with $p = 0.5$, is 1. For a random variable X with standard random variable Z, follows that $\mathrm{Kurt}[X] = \mathrm{E}[Z^4] \geq \mathrm{E}[(Z^2)]^2 = 1$ by setting Z^2 and the function x^2 on Jensen's inequality. The unique distribution for Z such that Z^2 to be a constant random variable is the symmetric standard two-point distribution, and therefore, the ratio is $\rho = 0.5$, as desired.

2.7 Conditional probability distribution

To introduce the discussion about conditional probability distributions, following the structure of Hoel, Port, and Stone [31], we start considering the discrete case. Let X and Y be discrete random variables. If we want to get the pmf of the random variable such that Y given that $X = x$, denoted $Y|X$, we get it directly from Def. 2. Thus,

$$f_{Y|X}(y|x) = \mathrm{P}[Y = y|X = x] = \frac{\mathrm{P}[X = x, Y = y]}{\mathrm{P}[X = x]} = \frac{f_{X,Y}(x, y)}{f_X(x)}, \tag{2.62}$$

for $x \in R_X$. On the other hand, if X is continuous, then $\mathrm{P}[X = x] = 0$ and therefore $\mathrm{P}[Y = y|X = x]$ is always undefined. To handle this case, we consider firstly the cdf.

Thus, for finite $x \in R_X$, we consider the concept of limit in such a way that

$$F_{Y|X}(y|x) = \mathrm{P}[Y \leq y|X = x] = \lim_{\Delta \to 0^+} \mathrm{P}[Y \leq y|x - \Delta \leq X \leq x + \Delta]. \tag{2.63}$$

It can be shown that the limit results in

$$F_{Y|X}(y|x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x,y)\ dy}{f_X(x)}. \tag{2.64}$$

Taking the derivative with respect to y gives the same as Eq. (2.62), leading to the following definition.

Definition 15 (Conditional probability distribution) *For random variables X and Y, the conditional probability distribution of Y given $X = x$, denoted $f_{Y|X}$ for an associated random variable $Y|X$, is given by*

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \tag{2.65}$$

with $R_{Y|X} = \{x : 0 < f_X(x) < \infty\}$.

As $f_{Y|X}(y|x)$ is a probability distribution like any other, everything that has been discussed about probability distribution and will be discussed from here on applies to conditional distributions. The expectation of $f_{Y|X}(y|x)$ is referred to as conditional expectation, the variance as conditional variance, and so on.

2.7.1 Truncated distribution

A particular case of conditional probability of much importance for this work is the truncated distributions. Let X be a random variable and $X_{(a,b]}$ be another random variable such that X given $a < X \leq b$. Follows from Def. 15 that

$$f_{X_{(a,b]}}(x) = \frac{f_X(x)}{F_X(b) - F_X(a)}, \tag{2.66}$$

with $R_{X_{(a,b]}} = \{x \in R_X : a < x \leq b\}$. The distribution $f_{X_{(a,b]}}(x)$ is known as truncated probability distribution or simply truncated distribution, since it restricts the support of the original probability distribution $f_X(x)$.

The conditional expectation can be calculated as

$$\mathrm{E}[X_{(a,b]}] = \frac{\sum\!\!\!\!\!\!\int_{x \in R_{X_{(a,b]}}} x f_X(x)}{F_X(b) - F_X(a)}. \tag{2.67}$$

In particular, if $X_{\leq x}$ and $X_{>x}$ are the random variables of X given $X \leq x$ and X

given $X > x$, respectively, we have

$$f_{X_{\leq x}}(k) = \frac{f_X(k)}{F_X(x)}, \; f_{X_{>x}}(k) = \frac{f_X(k)}{1 - F_X(x)}, \tag{2.68}$$

with $R_{X_{\leq x}} = \{k \in R_X : k \leq x\}$ and $R_{X_{>x}} = \{k \in R_X : k > x\}$, and

$$\mathrm{E}[X_{\leq x}] = \frac{\mathop{\sum\!\!\!\!\!\!\int}_{k \in R_X : k \leq x} k f_X(k)}{F_X(x)}, \; \mathrm{E}[X_{>x}] = \frac{\mathop{\sum\!\!\!\!\!\!\int}_{k \in R_X : k > x} k f_X(k)}{1 - F_X(x)}. \tag{2.69}$$

Denoting $\mathrm{E}[X_{\leq x}]$ and $\mathrm{E}[X_{>x}]$ by $\mathrm{E}[X|X \leq x]$ and $\mathrm{E}[X|X > x]$, respectively, and introducing the quantity

$$G_X(x) = \mathop{\sum\!\!\!\!\!\!\int}_{k \in R_X : k \leq x} k f_X(k), \tag{2.70}$$

we have

$$\mathrm{E}[X|X \leq x] = \frac{G_X(x)}{F_X(x)}, \; \mathrm{E}[X|X > x] = \frac{\mu_X - G_X(x)}{1 - F_X(x)}, \tag{2.71}$$

since

$$\mu_X = \mathop{\sum\!\!\!\!\!\!\int}_{x \in R_X} x f_X(x) = G_X(t) + \mathop{\sum\!\!\!\!\!\!\int}_{x \in R_X : x > t} x f_X(x). \tag{2.72}$$

A basic fact of $G_X(x)$ is that

$$\lim_{x \to R_X^{max}} G_X(x) = \mu_X. \tag{2.73}$$

Note that for continuous distributions $\frac{dG_X(x)}{dx} = x f_X(x)$, while for discrete ones we can write $G_X(x)$ by using the step function in an analog way to $F_X(x)$ as

$$G_X(x) = \sum_{k \in R_X} k f_X(k) \theta(x - k). \tag{2.74}$$

That way, we have $\frac{dG_X(x)}{dx} = x f_X^G(x)$.

Furthermore, the quantity $G_X(x)$ is affected by location shifting and change of the scale with $a \in \mathbb{R}$ and $b > 0$ by $G_{X+a}(x) = aF_X(x - a) + G_X(x - a)$ and $G_{bX}(x) = bG_X(x/b)$, respectively. For the discrete case, it follows from

$$\begin{aligned} G_{X+a}(x) &= \sum_{k \in R_{X+a} : k \leq x} k f_{X+a}(k) = \sum_{k \in R_X : k \leq x - a} (k + a) f_X(k) \\ &= aF_X(x - a) + G_X(x - a) \end{aligned} \tag{2.75}$$

and

$$G_{bX}(x) = \sum_{k \in R_{bX} : k \leq x} k f_{bX}(k) = \sum_{k \in R_X : k \leq x/b} bk f_X(k) = bG_X(x/b). \tag{2.76}$$

The continuous is analogous but by using substituting of variables on the integration.

To finish, the reflected effect on $G_X(x)$ is given by

$$\begin{aligned} G_{-X}(x) &= \sum\!\!\!\!\!\!\!\int_{k\in R_{-X}:k\leq x} k f_{-X}(k) = -\sum\!\!\!\!\!\!\!\int_{k\in R_X:k\geq -x} k f_X(k) \\ &= G_X(-x) - \mu_X + x\,\mathrm{P}[X=-x], \end{aligned} \tag{2.77}$$

where the last equality follows from Eq. (2.72).

2.8 Characteristic function

Characteristic functions (CF) are the Fourier transforms of probability distributions, an example of the concept of Probability Theory brought from other branches of mathematics [31]. These quantities are a convenient way to represent distributions in many situations since there is a one-to-one correspondence between cumulative distribution functions and characteristic functions, a result known as the uniqueness theorem. The characteristic functions are defined as follows.

Definition 16 (Characteristic function) *The characteristic function of a random variable X, denoted $\varphi_X(\omega)$, is given by $\varphi_X(\omega) = \mathrm{E}[e^{i\omega X}]$ for $-\infty < \omega < \infty$.*

Note that $e^{i\omega X}$ is a complex-valued random variable. The definition of complex random variables is analog to the real-valued ones. A complex random variable Z can be written with its real and imaginary components in such a way that $Z = X + iY$, where X and Y are real-valued random variables. Provided that $\mathrm{E}[X]$ and $\mathrm{E}[Y]$ are well-defined, the expected value of Z is defined as

$$\mathrm{E}[Z] = \mathrm{E}[X+iY] = \mathrm{E}[X] + i\,\mathrm{E}[Y]. \tag{2.78}$$

Such as on real-valued random variables, Z has finite expectation if $\mathrm{E}[|Z|] \leq \infty$. Since, by linearly, LOTUS can be extended to complex-valued function, we get

$$\varphi_X(\omega) = \sum\!\!\!\!\!\!\!\int_{x\in R_X} f_X(x) e^{i\omega x}, \tag{2.79}$$

which means that the characteristic function is equivalent to continuous-time Fourier transform in continuous case and to discrete-time Fourier transform in discrete case [57], sharing their properties.

There is a quantity closed related to characteristic functions, the moment generating functions (MGF), denoted $M_X(\omega)$ and defined as $M_X(\omega) = \mathrm{E}[e^{\omega X}]$. The domain of $M_X(\omega)$ is restricted to the values of ω in which $e^{\omega X}$ has finite expectation. With LOTUS, $M_X(\omega) = \sum\!\!\!\!\!\!\int_{x\in R_X} f_X(x)e^{\omega x}$. Note that MGF for continuous random variables is equivalent to bilateral Laplace transform, while for discrete is equivalent to bilateral Z-transform up to a change on the variable [57]. It can be shown that if $M_X(\omega)$ is finite on $-\omega_0 \leq \omega \leq \omega_0$ for some positive ω_0, then all moments of X are finite. In that case, the expansion

$$M_X(\omega) = \sum_{n=0}^{\infty} \frac{\mathrm{E}[X^n]}{n!}\omega^n \tag{2.80}$$

holds for $-\omega_0 \leq \omega \leq \omega_0$. Comparing the coefficients of Eq. (2.80) with the coefficients of the Taylor expansion of the own $M_X(\omega)$, we can get an expression to calculate the nth raw moment with

$$\mathrm{E}[X^n] = \frac{d^n}{d\omega^n} M_X(\omega)\bigg|_{\omega=0}, \tag{2.81}$$

justifying the terminology of the moment generating function.

We can apply it also to characteristic functions. If X has the nth moment, the nth derivative of $\varphi_X(\omega)$ is given by

$$\frac{d\varphi_X(\omega)}{d\omega} = i^n\,\mathrm{E}[e^{i\omega X} X^n], \tag{2.82}$$

which given in the particular of $\omega = 0$,

$$\mathrm{E}[X^n] = i^{-n}\frac{d}{d\omega}\varphi_X(\omega)\bigg|_{\omega=0}. \tag{2.83}$$

Furthermore, provided that the expansion of Eq. (2.80) holds in the interval for $-\omega_0 \leq \omega \leq \omega_0$, the expansion

$$\varphi_X(\omega) = 1 + \sum_{n=1}^{\infty} \frac{i^n\,\mathrm{E}[X^n]}{n!}\omega^n \tag{2.84}$$

also holds in the same interval.

Characteristic functions are a complex extension of moment generating functions, consequently containing convenient algebraic properties that allow, for instance, to prove the CLT [31]. Another important vantage of characteristic functions is to be finite for all real numbers ω. The reason is because $|e^{i\omega x}|$ is bounded. More strongly, the characteristic function has the property

$$|\varphi_X(\omega)| = |\,\mathrm{E}[e^{i\omega X}]| \leq \mathrm{E}[|e^{i\omega X}|] = 1. \tag{2.85}$$

In particular, $\varphi_X(0) = 1$. In general, assuming that the nth moment exist, the nth derivative can be bounded from Eq. (2.82) as

$$\left|\frac{d^n \varphi_X(\omega)}{d\omega^n}\right| \leq \mathrm{E}[|X|^n]. \tag{2.86}$$

Now, recall that the sum of i.i.d. random variables $X_1, \dots, X_n$ is a sequence of convolutions. Thus, we know from Fourier transform theory that

$$\varphi_{X_1+\ldots+X_n}(\omega) = \varphi_{X_1}(\omega)\ldots\varphi_{X_n}(\omega). \tag{2.87}$$

As the characteristic function admits an inverse function (that can be founded, for instance, in Hoel, Port, and Stone [31]), the property of Eq. (2.87) is a convenient manner of computing the sum of the i.i.d. random variables. Other important properties of the characteristic function, some of them brought from known results of Fourier transform theory, are listed.

- A characteristic function $\varphi_X(\omega)$ is a continuous function of ω;
- **(Location shifting)** For any $a \in \mathbb{R}$, $\varphi_{X+a}(\omega) = e^{i\omega a}\varphi_X(\omega)$;
- **(Change of the scale)** For any $b > 0$, $\varphi_{bX}(\omega) = \varphi_X(b\omega)$;
- **(Reflected random variable/Hermiticity)** $\varphi_{-X}(\omega) = \varphi_X(-\omega) = \varphi^*_X(\omega)$[11];
- **(Symmetric distributions)**[12] If $f_X(x)$ is a symmetric distribution (even function), then $\varphi_X(\omega)$ is real and even.

To finish, let's give attention to the first derivative of the characteristic function, an important quantity for the results of this work. From Eq. (2.82),

$$\varphi'_X(\omega) = i \sum\!\!\!\!\!\!\int_{x\in R_X} x f_X(x) e^{i\omega x}. \tag{2.88}$$

For $a \in \mathbb{R}$ and $b > 0$,

$$\varphi'_{X+a}(\omega) = \frac{d}{d\omega}[e^{i\omega a}\varphi_X(\omega)] = e^{ia\omega}(ia\varphi_X(\omega) + \varphi'_X(\omega)) \tag{2.89}$$

and a changing of variable gives $\varphi'_{bX}(\omega) = b\varphi'_X(b\omega)$. Furthermore, $\varphi'_{-X}(\omega) = \frac{d}{d\omega}\varphi_X(-\omega) = -\varphi'_X(-\omega)$.

[11] The symbol z^* on a complex number z denotes the complex conjugate of z.

[12] Of course, we can ignore the analog property for odd functions on probability distributions.

Chapter 3

The Quantum Alternating Operator Ansatz and the Grover Mixer

This chapter has the main objective of presenting the two variants of Quantum Alternating Operator Ansatz that are the main object of study of this work, the GM-QAOA, and GM-Th-QAOA. To get that, we go through the important topics of combinatorial optimization, Grover's algorithm, VQAs, QAA, and QAOA (both the Quantum Approximate Optimization Algorithm and the Quantum Alternating Operator Ansatz). We also present the QWOA, in which the particular case of the complete graph can be seen as an alternative formulation of the Grover mixer; and the algorithms GAS and MAOA, both relevant to the discussion.

We assume that the reader has knowledge of quantum computing. If this is not the case, we recommend for an extensive study the traditional books of Nielsen and Chuang [1] and Kaye, Laflamme, and Mosca [58] or for a more introductory study the books of Marquezino, Portugal, and Lavor [59], and Portugal [60]. In addition to the articles cited throughout the chapter, we use the Nielsen and Chuang [1] book to the geometric interpretation of Grover's algorithm, and for some sections, the review papers Cerezo et al. [4] and Blekos et al. [10] of VQA and QAOA, respectively. In particular, to the discussion of combinatorial optimization on Sec. 3.1, we use the books of Cormen et al. [61], Szwarcfiter [62], Bernhard and Vygen [9], and Skiena [63]. We also assume a basic knowledge of graph theory, such as in the concepts of vertex cover and the Hamiltonian cycle, recommending the Bondy and Murty [64] book, and on (classical) algorithms and complexity of algorithms, recommending the book of Cormen et al. [61].

3.1 Combinatorial optimization

Combinatorial optimization is an area of discrete mathematics that deals with optimization problems in finite sets. It is related to several other branches, such as combinatorics, operations research, graph theory, and theoretical computer science. Although problems of this nature date back to much older times, combinatorial optimization became an independent field only in the middle of the last century, and since then, countless real-world applications can be formulated as abstract combinatorial optimization problems [9].

A combinatorial optimization problem (COP) consists of finding the best object, called optimal object (or global optima), among a finite set S of discrete objects evaluated through the extremization of a real-valued function $c(k) : S \to \mathbb{R}$ called the objective function. The objects are also called solutions or feasible solutions. In the present work, we assume that the objective function must be minimized. It is straightforward to convert a maximization problem into a minimization one by multiplying the cost function by -1. Any algorithm defined from here for minimization, such as the Quantum Alternating Operator Ansatz, could be equivalently defined for maximization problems. The set of combinatorial objects S is called also the combinatorial domain, the set of feasible solutions, or the solution space.

For this work, we assume that the discrete objects are labeled by n-bit strings such that $S \subseteq \{0,1\}^n$. Following the widely used terminology in QAOA literature, we say a combinatorial optimization problem is unconstrained if $S = \{0,1\}^n$ and constrained otherwise. That classification is natural in the context of the gate model of quantum computing since we usually codify the combinatorial domain using qubits. Some constrained problems can be defined into known families of combinatorial objects, such as all permutations of n elements and all k-combinations of n elements [12].

3.1.1 Combinatorial optimization problems

In this subsection, we present the combinatorial optimization problems cited during this work. Since most of them are graph problems, we standardized the notation to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ representing a graph $\mathcal{G}$ with a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}$. We also denote an edge with extremes on vertices u and v by (u, v). Some classes of graphs have their own symbol, such as K_n for a complete graph with n vertices and $K_{j,k}$ for the complete bipartite graph with partitions of j and k vertices. For each COP listed, we present the input of the problem, the objective that must be codified on the objective function, and the combinatorial domain. A specific input is known as an instance of the problem.

- Max-Cut [10].
  - Input: a graph $\mathcal{G}$.
  - Objective: find a partition of the set of vertices $\mathcal{V}$ into two complementary subsets that maximize the number of edges between both subsets, called cut edges.
  - Combinatorial domain: unconstrained problem of $|\mathcal{V}|$ bits.
- Max k-SAT [17].
  - Input: n boolean variables organized into m clauses of length k[1].
  - Objective: find a variable assignment that satisfies the maximum number of clauses.
  - Combinatorial domain: unconstrained problem of n bits.
- Number Partition Problem [30].
  - Input: set a positive real numbers $\{x_0, \dots, x_n\}$.
  - Objective: find a partition of the set $\{x_0, \dots, x_n\}$ into two complementary subsets that minimize the difference between the sum of both subsets.
  - Combinatorial domain: unconstrained problem of n bits.
- Traveling Salesman Problem (or Traveling Salesperson Problem) [13].
  - Input: a complete graph K_n in which each edge of $\mathcal{E}$ is associated with a non-negative real number called weight (i.e., a weighted graph).
  - Objective: find a Hamiltonian cycle with the lowest sum of the weights of their edges.
  - Combinatorial domain: some bit string codifying the permutations of n elements.
- Minimum Vertex Cover [14].
  - Input: a graph $\mathcal{G}$.
  - Objective: find a vertex cover with the minimum number of vertices.
  - Combinatorial domain: until the author's knowledge, not identifiable with a known combinatorial family in general graphs.
- Max k-Vertex Cover [11].

[1] Here, a clause with length k is a disjunction of k literals. A literal is either a boolean variable or the negation of a boolean variable.

  - Input: a graph $\mathcal{G}$.
  - Objective: find a set of k vertices that cover the largest number of edges.
  - Combinatorial domain: k-combinations of $|\mathcal{V}|$ bits. Alternatively, we can say that the combinatorial domain consists of the $|\mathcal{V}|$-bit strings with Hamming weight equal to k.

- k-Densest Subgraph [11].
  - Input: a graph $\mathcal{G}$.
  - Objective: find a subgraph with k vertices that have the largest number of edges.
  - Combinatorial domain: k-combinations of $|\mathcal{V}|$ bits/$|\mathcal{V}|$-bit strings with Hamming weight equal to k.

- Max Bisection [11].
  - Input: a graph $\mathcal{G}$ with an even number of vertices.
  - Objective: find a partition of $\mathcal{G}$ into two subgraphs with the same number of vertices that maximize the edges between the partitions—note that Max-Cut with the restriction of the partitions has the same size becomes Max Bisection.
  - Combinatorial domain: k-combinations of $|\mathcal{V}|$ bits/$|\mathcal{V}|$-bit strings with Hamming weight equal to k.

The problem of Capacitated Vehicle Routing [24] and the problems based on Portfolio Optimization [13, 25, 40] have more elaborate definitions that can be found in their own papers.

3.1.2 Decisions problems and complexity classes

The class of combinatorial optimization problems belongs to the more general class of optimization problems, which can be defined as the class of problems in which we must find a structure that satisfies certain optimization criteria. Two other classes of problems are search problems and decision problems. The first one consists of the problems in which we have to find a structure that satisfies a property, such as the unstructured search problem, presented in the next section. In the last one, we do not want to find the structure that satisfies a given property, but only decide if this structure exists. Decision problems are questions of yes or no answers. An example of a decision problem is the illustrious Satisfiability problem or SAT problem, the first problem to be proved as NP-Complete [65][2]. The SAT problem asks if there is an assignment of the boolean variables of a given set of clauses such that all clauses are satisfied. Note that the Max k-SAT problem, presented in the previous subsection, is an optimization variation of SAT.

Historically, the theory of computational complexity was developed concerning decision problems [65, 66], with that well-known class of complexity P, NP, NP-Hard, NP-Complete, in which the definitions can be founded, for instance, in Cormen et al. [61] book, being directly classifications for decision problems. Although the issue of NP-Completeness for optimization problems in general case is something more delicate [67], in this work, we say that an optimization problem belongs to one of these classes if its associated decision problem also belongs. That is because an optimization problem can be immediately converted into a decision problem. For a given value k, we can ask if there is a solution such that the objective function outputs a value equal/larger or equal/smaller than k. For instance, for the Max-Cut problem, we can consider that problem that asks for given a graph $\mathcal{G}$ and a positive integer k if there is a partition of the set of vertices $\mathcal{V}$ into two complementary subsets such that number of edges between both subsets is at least k.

3.1.3 Classical approaches

Usually, the combinatorial optimization problems of interest are NP-Hard—as the case of the problems presented on Subsec. 3.1.1. To these problems, there is no efficient algorithm for exact optimization (or global optimization) unless P=NP. In that case, we can only resort to the pursuit of algorithms that output approximate solutions. In the terminology of this work, we split the algorithms that do not aim the exact optimization into approximation algorithms and heuristics. The difference between them is that only approximation algorithms provide some guarantee based on the metric that relates the output solution to the minimum one. In the majority of cases in the literature, the metric used as a performance guarantee is the approximation ratio (or approximation factor) λ, defined if the minimum solution is non-zero as

$$\lambda = \frac{c_{out}}{c_{min}}, \tag{3.1}$$

where c_{out} is the output solution and c_{opt} minimum one. Although it is not orthodox, for technical purposes, we allow in this work the approximation ratio to be negative in situations when the objective function can output both positive and negative values. For instance, if $c_{out} = 2$ and $c_{min} = -4$, then $\lambda = -0.5$. The approximation ratio of a maximization problem with non-negative values, such as Max-Cut, is not affected by the conversion on a minimization problem. Follows that if $c_{min} < 0$, then

[2] Although the concept of NP-Completeness was later introduced by Karp [66].

$\lambda \le 1$ and if $c_{min} > 0$, then $c_{min} \ge 1$.

In classical computing, examples of approximation algorithms for specific problems are the Christofides algorithm [9] for Traveling Salesman Problem, which guarantees $\lambda = 3/2$ for metric instances, i.e., in which the weights of edges obey the triangle inequality; and the Goemans-Williamson algorithm [68], which gives a guarantee of $\lambda \approx 0.8786$ on Max-Cut problem. In particular, unless P=NP, there is no efficient algorithm to approximate the Traveling Salesman Problem with general instances on a constant approximation ratio [9]; and approximate Max-Cut beyond the approximate ratio $\lambda = 16/17 \approx 0.9412$ in general case is proved to be NP-Hard [69, 70].

On the other hand, there are classical approaches used for general combinatorial optimization problems. The most rudimentary of all is the classical brute force approach. To the exact optimization, the classical brute force consists of simply verifying all possible feasible solutions, which has the runtime $\Theta(|S|)$ if we do not know how to identify when a solution is optimal. On non-exact optimization, we can consider the called classical random sampling (CRS) [24], in which we uniformly sampling a given number of times the solution space and take the smallest one. If we model the random sample of the solution space with random variables[3], CRS is equivalent to the first order statistic. One refinement of the classical brute force to exact optimization approach is called backtracking. It refers to a general technique in which the solution space is systematically enumerated, allowing sets of solutions can be discarded without explicit consultation by detecting some certificate indicating these solutions are not optimal.

A heuristic still very simple is the Hill Climbing [71], which starts from a given solution and evaluates the objective function for some defined neighboring solutions. We pick the best neighbor and repeat the process until finding a solution better than all its neighbors. The main problem with this approach is that it can fall into so-called local minima, which may correspond to unsatisfactory solutions. More robust heuristics are simulated annealing, an algorithm inspired by the cooling of physical systems, and genetic algorithms, which are a class of algorithms inspired by the process of natural selection of the Theory of Evolution. Furthermore, several COPs can be formulated as integer programming problems, which have their own arsenal of algorithms [72].

In the quantum computing case, it is safe to say that the most prominent general heuristic for combinatorial optimization is the QAOA, discussed further. The state-of-art for quantum computing for combinatorial optimization (and optimization in general) can be founded on the review paper of Abbas et al. [73].

[3] See Chapter 4 to a more precise definition of the model of the solution space with random variables.

3.2 Grover's algorithm and the unstructured search problem

The unstructured search problem can be stated as follows. Consider the function

$$f(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise,} \end{cases} \tag{3.2}$$

with a finite domain of size N and where A is a subset of that domain of size n. Suppose that the elements of the A are unknown, and we want to find it with as few evaluations of any point of the domain as possible. The function $f(x)$ is called an oracle, and the elements of A are known as marked elements[4]. We denote the ratio n/N of marked elements on the entire domain by ρ. The complexity in terms of the number of calls of the oracle is called query complexity. Although the unstructured search problem is essentially a search problem, it can be formulated as an optimization problem, for instance, for the context of the VQAs [74, 75].

Of course, we are not interested in best-case complexity since, in classical computing, we could find a marked element on the first evaluation if it happens to be in the first position consulted. In the quantum computing case, we could build an algorithm that only amplifies the probability of specific states that, by chance, could be marked in a given instance. Therefore, it is reasonable to consider both average and worst cases. For this work, we are considering the average-case complexity. Respecting our intuition, the best classical algorithm for unstructured search problems on the average-case is a classical brute force evaluation of the oracle to distinct points on the domain until finding a marked element. In that case, the average-case query complexity is $\Theta(1/\rho)$ as $\rho \to 0$.

In the quantum computing case, the situation becomes more interesting due to the possibility of simultaneous evaluations of different points of domains. Combining it with an intelligent exploration of the quantum interference phenomenon, we can go beyond the classical brute force. In this sense, quantum computing has Grover's algorithm, also called Grover's search, originally introduced to the unstructured search problem with a single marked element, i.e., $n = 1$, and with N being a power of 2 [18, 19]—to encode the points of the domain precisely on $\log_2(N)$ qubits—and later generalized to the version presented here [76]. Grover's algorithm has a query complexity of $\Theta(1/\sqrt{\rho})$ as $\rho \to 0$, a quadratic gain over classical brute force.

The evaluation of function $f(x)$ is given on quantum computing by the called quantum oracle O, which is a black box unitary that is somehow able to distinguish

[4] The marked/non-marked elements also can be called good/bad elements or elements as winning/losing.

between marked and non-marked elements. In particular, the oracle of Grover's algorithm, also called phase oracle, denoted O_G, acts in a state $|x\rangle$ as

$$O_G|x\rangle = \begin{cases} -|x\rangle, & x \in A \\ |x\rangle, & \text{otherwise} \end{cases} = (-1)^{f(x)}|x\rangle. \tag{3.3}$$

Grover's algorithm can be present also with the standard oracle O_S, which uses an ancilla qubit and acts as $O_S|x\rangle|j\rangle = |x\rangle|j \oplus f(x)\rangle$, where $\oplus$ is the binary sum or bitwise xor. In this work, we consider only the phase oracle.

On a quantum algorithm for the unstructured search problem, the alone oracle's action is not enough to measure the marked elements with high probability. We need an operator to generate interference between states with the goal of amplifying the probability of measuring the marked states. That type of operator is usually called the diffusion operator. In Grover's algorithm, the diffusion operator, denoted by D_G, is known as Grover's diffusion operator and is given by

$$D_G = 2|d\rangle\langle d| - \mathbb{I}. \tag{3.4}$$

The symbol $\mathbb{I}$ denotes the identity matrix, and $|d\rangle$ is a uniform superposition over all states of the computational basis codifying points of the domain of $f(x)$, that is,

$$|d\rangle = \frac{1}{\sqrt{M}} \sum_{x \in \text{Dom}\{f\}} |x\rangle. \tag{3.5}$$

Denoting the combining application of both operators by $G = D_G O_G$, the final state of the unitary evolution of Grover's algorithm, denoted $|\psi^{(r)}\rangle$, consists in a number of r of applications of the operator G on the initial uniform superposition $|d\rangle$. A single application of G is called Grover's iteration or Grover's round. The number of rounds to measure a marked state with high probability from the state $|\psi^{(r)}\rangle$ is of order $\Theta(1/\sqrt{\rho})$ as $\rho \to 0$.

The compilation of Grover's diffusion operator on a quantum circuit can be done with time complexity $\mathcal{O}(\log(N))$ in terms of universal gates. Combining it with the query complexity gives a general runtime of $\mathcal{O}(1/\sqrt{\rho}\log(N))$. The space complexity is in order of $\mathcal{O}(\log(N))$. The procedure is discussed by Portugal [22] for standard oracle formulation and N as a power of 2. However, we can convert in the phase oracle formulation with the procedure of Portugal [60], and for general N, as suggested by Boyer et al. [76], we can use the approximate Fourier transform given by Kitaev [77].

3.2.1 Geometric interpretation of Grover's algorithm

For the purposes of this work, we ignore the quantum circuit compilation of Grover's algorithm and focus only on the analysis of the unitary dynamics of the algorithm. Since Grover's algorithm acts on the subspace of real numbers and all marked/non-marked elements share the same amplitudes, it admits a geometric interpretation in a two-dimensional reduced subspace.

Denoting by $|a\rangle$ and $|b\rangle$ uniform superpositions over all non-marked elements and marked elements, respectively, we can write the initial state as

$$|d\rangle = \cos(\theta/2)|a\rangle + \sin(\theta/2)|b\rangle \tag{3.6}$$

where $\theta/2 = \arcsin(\rho)$ is angle of the vector $|d\rangle$ with the axis $|a\rangle$. The application of the oracle operator is a reflection over the vector $|a\rangle$ while the diffusion operator acts as a reflection on the own $|d\rangle$ vector. The combined action of both increases the angle of the state vector to $3\theta/2$ radians. Fig. 3.1 illustrates the geometrical interpretation for a single iteration.

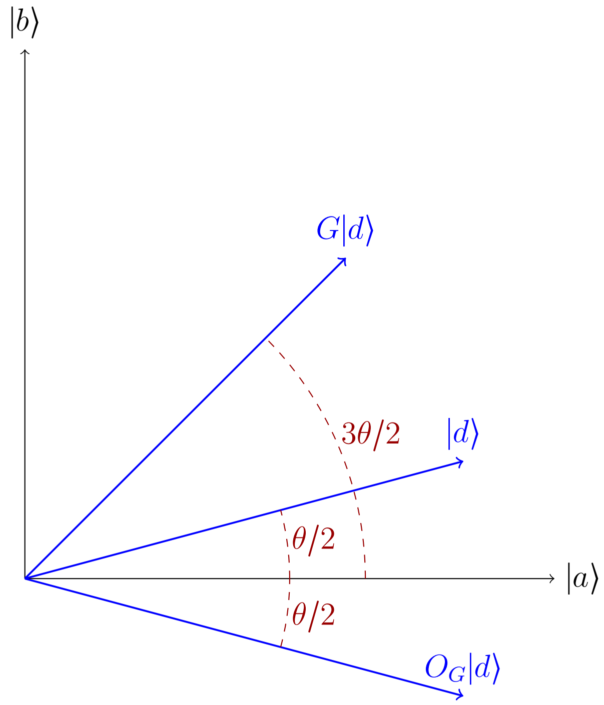


Figure 3.1: Geometric interpretation of a single Grover's iteration. Two-dimensional subspace spanned by the vectors $|b\rangle$ and $|a\rangle$. The initial state $|d\rangle$ makes an angle of $\theta/2$ with the axis $|a\rangle$. The application of the oracle and diffusion operator are reflections over $|a\rangle$ and $|d\rangle$, respectively, which results in an angle of $3\theta/2$ for $G|d\rangle$. The creation of this figure by the author was inspired by Figure 6.3 of Nielsen and Chuang [1].

In general, one can show that each iteration increases the angle of the state vector in θ radians. Therefore,

$$|\psi^{(r)}\rangle = G^r|d\rangle = \cos\left((2r+1)\theta/2\right)|a\rangle + \sin\left((2r+1)\theta/2\right)|b\rangle. \tag{3.7}$$

Denoting the event "suc" as the success of the algorithm, that is, the measurement of a marked element, we have

$$\text{P[suc]} = \sin^2\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right). \tag{3.8}$$

In particular, if the angle of geometric interpretation $(2r+1)\theta/2$ is $\pi/2$ radians, then P[suc] = 1. Solving it gives the runtime $r_{opt} = \left\lfloor \frac{\pi}{4\sqrt{\rho}} \right\rfloor$. One can verify that, for instance, $\rho = 0.25$ gives $r_{opt} = 1$ with exact P[suc] = 1. Of course, the probability of being equal to 1 does not hold in general. However, the optimal probability is bounded by P[suc] $\geq 1 - \rho$, being therefore 1 asymptotically on $\rho \to 0$. Furthermore, note that since the position of the marked elements on the domain is irrelevant to the performance of the algorithm, the complexity of worst, best, and average cases are the same.

3.2.2 The low-convergence regime and the maximum amplification

Taking $\rho \to 0$ and $\rho << 1/(2r+1)^2$ in Eq. (3.8) reduce the probability P[suc] of Grover's algorithm to $\rho(2r+1)^2$ since $\sin(x) \to x$ and $\arcsin(x) \to x$. Bennett and Wang [23] refer to it by the low-convergence regime of Grover's search.

Until the angle of $\pi/2$ radians on a geometric interpretation of Grover's algorithm, we denote the ratio P[suc]$/\rho$ by η. That represents the ratio of the probability of measuring a marked state before and after the application of Grover's iterations, i.e., the amplification of probability realized by Grover's algorithm. The low-convergence regime maximizes the amplification with $(2r+1)^2$. We prove it in Appendix A, showing additionally that for a fixed r, η is strictly decreasing in the function of ρ. Fig. 3.2 illustrates how small values of ρ get closer to the low-convergence regime for $r = 1$. As noticed by Bennett and Wang [23], the accuracy between the low-convergence regime and the true amplification is within 1% when the amplified probability is less than 1/40.

3.2.3 Optimality of Grover's algorithm on the unstructured search problem

A natural question is whether Grover's algorithm is optimal on average-case for the unstructured search problem on quantum computing. The answer is yes. The

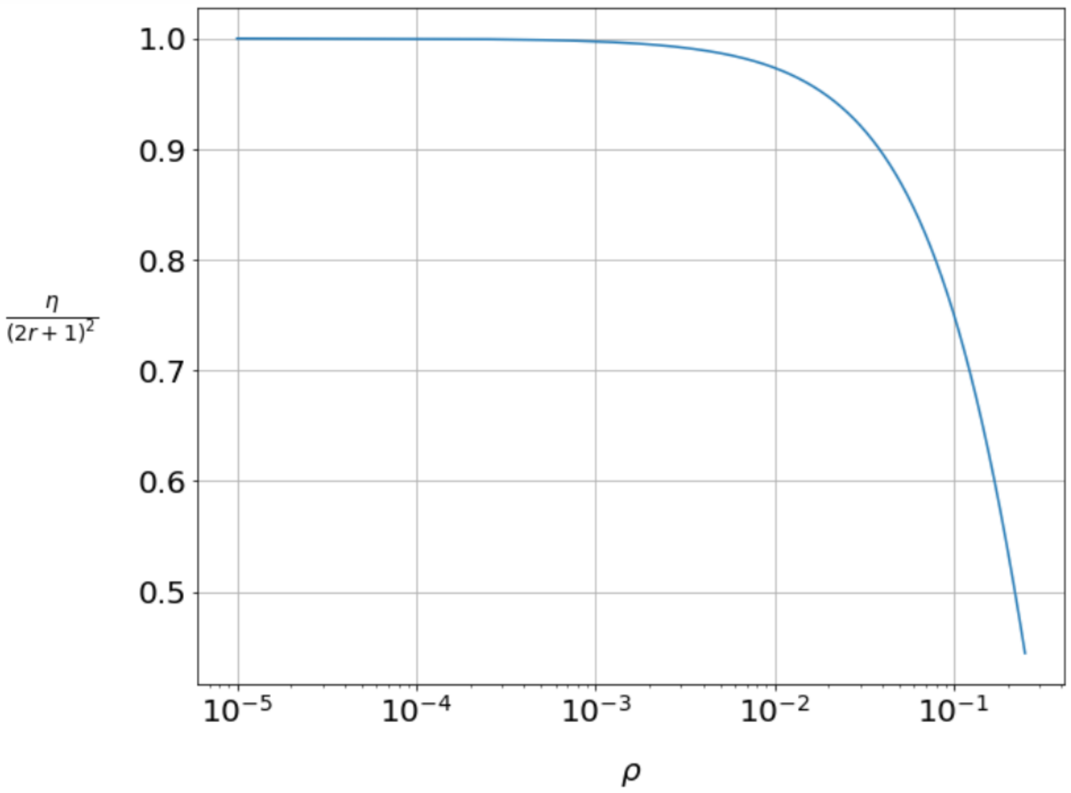


Figure 3.2: Amplification η versus the ratio ρ for the range 10^{-5} until 0.25 on linear-log scale and $r = 1$. The ratio $r = 0.25$, which has an amplification of $\eta = 4$, gives the $\pi/2$ radians of the geometric interpretation of Grover's algorithm. We normalize η by the maximum amplification of $(2r + 1)^2$. For low ratios, the amplification gets closer to the maximum amplification, while for ratios near to $\pi/2$ radians on geometric interpretation, η is considerably lower than $(2r + 1)^2$.

first result concerning the optimality of Grover's algorithm was present in Bennett et al. [29] paper, which proves that for $n = 1$ case, the number of oracle calls of a quantum algorithm must be $\Omega(\sqrt{N})$, establishing the asymptotic optimality of Grover's algorithm. Subsequently, Boyer et al. [76] consider arbitrary n and show that Grover's search is at least near in a factor of 2 to be optimal in terms of the number of rounds. Zalka [35] went further and proved the strongest possible result for $n = 1$ case: Grover's algorithm is exactly optimal for the probability until the threshold angle of $\pi/2$ radians on the geometric interpretation—from which the probability starts to decrease. Finally, Hamann, Dunjko, and Wölk [36] generalize Zalka's proof for arbitrary n. Solving $\mathrm{P}[\mathrm{suc}] = \pi/2$ for ρ, we can conclude that the ratio point of $\pi/2$ radians is $\sin^2(\pi/(4r+2))$. We enunciate the result of Hamann, Dunjko, and Wölk [36] adapting to our notation on the following theorem.

Theorem 6 (Optimality of Grover's algorithm) *For an oracle O that marks exact n over N elements and such $\rho = n/N$, Grover's algorithm gives the maximum possible average probability of measuring a marked element for a quantum algorithm,* *given by Eq.* (3.8) *for up to the interval* $\rho \leq \sin^2(\pi/(4r+2))$.

3.3 Grover Adaptive Search

The insights provided by Grover's algorithm go far beyond the unstructured search problem. The essence of the algorithm is a technique called amplitude amplification [78] that is used in several quantum algorithms. For instance, a more general version of the amplitude amplification can be used to improve the runtime complexity of the Harrow–Hassidim–Lloyd (HHL) algorithm [79] for the problem of linear system of equations [80]. Beyond amplitude amplification, Grover's algorithm can used as a subroutine on algorithms for combinatorial optimization with the target of providing a quadratic speed-up over classical brute force [40]. An illustrious algorithm among them is the Grover Adaptive Search (GAS).

The origin of GAS goes back to the unstructured search problem for arbitrary n when the value of n is unknown. Note that the algorithm presented in the previous section started from the premise that the value of n is known. Otherwise, we would not know the required number of Grover's iterations. On the other hand, if the value of n is not known, we can apply a procedure called exponential quantum search, introduced by Boyer et al. [76]. Exponential quantum search is based on the classical algorithm of exponential search [81, 82], and the expected runtime is kept to the order $\mathcal{O}(1/\sqrt{\rho})$. The procedure is the following.

- Set $l = 1$ and $1 < K < 4/3$;
- Choose j uniformly at random among the elements of the set $\{0, 1, \ldots, \lceil l-1 \rceil\}$;
- Apply j Grover's iterations and denote the outcome as k;
- If k is a marked element, the procedure is over. Otherwise, update l to $\min\{Kl, \sqrt{N}\}$ and back to the second item.

That method is the heart of the minimization algorithm introduced by Durr and Hoyer [37], which can be seen as a variant of GAS, introduced in the papers [38, 39]. Grover Adaptive Search is a quantum computing implementation of a stochastic algorithm for global optimization called Hesitant Adaptive Search (HAS) [83].

Grover's algorithm is used iteratively in the GAS framework. In particular, the best value known so far is chosen as a threshold in which all values smaller than it are the marked elements of Grover's search. For exact optimization, GAS finds a minimum value of a given COP with probability at least 1/2 with a runtime of $\mathcal{O}(\sqrt{|S|})$, quadratic speed-up over the classical brute force algorithm. In general, when a quantum algorithm performs quadratically better than the classical brute force on some metric, we called a Grover-style speed-up or Grover-like speed-up.

Considering the Gilliam, Woerner, and Gonciulea [40] description, the procedure of GAS is following: the input is a $K > 1$ and a COP with combinatorial domain S and objective function $c(k)$. The algorithm begins uniformly sampling $k_1 \in S$ and setting $y_1 = c(k_1)$, $l = 1$, $j = 1$. Then, we repeat the following steps until a termination condition is met.

- Choose uniformly r_j from the set $\{0, 1, \ldots, \lceil l - 1 \rceil\}$;
- Apply the Grover's algorithm of r_j iterations with an oracle that marks all states $k \in S$ such that $c(k) < y_j$. The output solution is denoted k, while $y = c(k)$;
- If $y < y_j$ then $k_{j+1} = k$, $y_{j+1} = y$. and $l = 1$. Otherwise, $k_{j+1} = k_j$, $y_{j+1} = y_j$ and updated l to Kl;
- Update j to $j + 1$.

The termination condition can be on the number of repetitions of the steps, the runtime, or even another metric. Although the usual goal is the exact optimization for a runtime of $\mathcal{O}(\sqrt{|S|})$, one can use it as a heuristic algorithm with a more limiting termination condition.

The greatest challenge in the implementation of GAS is the compilation of the subroutine of Grover's algorithm, a topic discussed in Gilliam, Woerner, and Gonciulea [40] paper. In particular, an efficient implementation of Grover's oracle in general can be done with quantum arithmetic. However, this approach can become costly due to the Toffoli gates, making it prohibitive for NISQ devices. For this context, the aforementioned work presents a more appropriate method to build oracles for the class of problems called Constrained Polynomial Binary Optimization (CPBO), which generalizes the well-known Quadratic Unconstrained Binary Optimization (QUBO) problems, applying that framework to the Portfolio Optimization problem.

3.4 Variational Quantum Algorithms

The Variational Quantum Algorithms (VQA) [4] are a prominent class of optimization algorithms used to target optimization problems in the NISQ era. The optimization process is only possible because of the variational principle of quantum mechanics, which states that the expectation value of an observable for a given trial wave function is always equal to or greater than the ground state energy. With the variational principle, we can optimize the expectation value varying the trial wave function upon an ansatz[5] until finding the ground state. On quantum circuit model, the expectation value of a Hamiltonian H on a state $|\psi\rangle$, given by $\langle\psi|H|\psi\rangle$, is expected value of energy (eigenvalues) spectrum of H weighted by the probability of the states on $|\psi\rangle$. Mathematically, the variational principle states

$$\lambda_{min} \leq \langle\psi|H|\psi\rangle, \tag{3.9}$$

where λ_{min} is the lowest eigenvalue of H. Eq. (3.9) holds for Hermitian operators, which is the case of Hamiltonian operators.

VQAs are algorithms that combine quantum and classical computing through a hybrid loop, also called an outer loop or variational loop. On the quantum part, the ansatz, we apply on an initial state a parameterized unitary transformation $U(\boldsymbol{\theta})$ for some set of discrete or continuous (or both) optimization parameters $\boldsymbol{\theta}$ called variational parameters and then we make measurements. By a given number of measurements of the quantum circuit, we get a statistical estimator of a set of observables $\{O_k\}$. Classically, based on the quantum experiments, we update the optimization parameters using an optimizer with the goal of minimizing a real-valued function $c(\boldsymbol{\theta})$ called cost function that depends on the observables $\{O_k\}$ given by measurements. When applied to COPs, it is straightforward to consider the objective function as the cost function. The cost function induces a hypersurface called the cost landscape such that the task of the classical optimizer is to navigate through the landscape until finds a global minima, in a process also called training. We can think of VQAs as the quantum analog of classical machine learning methods, such as neural networks [4].

3.4.1 Ansatzes

As discussed in detail by Cerezo et al. [4], there are different classifications for the ansatzes. If the architecture used on the ansatz depends on the task that we are dealing with, we called problem-inspired ansatz. An example is the Unitary Coupled Clustered (UCC) [85] ansatz, used on quantum chemistry problems. In contrast, there are generic ansatz architectures called problem-agnostic, used independent of the availability of information on the problem. In some sense, the Quantum Alternating Operator Ansatz, present in Sec. 3.7, is an example of a problem-agnostic ansatz.

Another classification is the variable structure ansatz, which beyond the usual optimization of variational parameters on quantum gates of a fixed structure, optimizes the structure itself, adding and removing parts of the circuit—that propose

[5] In a general sense, an ansatz in physics and mathematics is an educated guess for the functional form of a solution of an equation or other problem [84].

of ansatz is introduced in a framework called ADAPT-VQE [86]; and the hardware efficient ansatz, that as the terminology suggests, are designed to reduce the depth on specific quantum hardware—an example is found on Kandala et al. [87] paper.

3.4.2 Optimizers and the barren plateau phenomena

The training of VQAs is a huge challenge since globally optimizing the classical optimization problems associated with VQAs is expected to be NP-Hard in many cases [88]. That issue, added to other challenges such as the barren plateau (BP) [89] problem—discussed further—underscore the importance and the necessity of the study of classical optimizers in the VQA's research. Following the classification of Cerezo et al. [4], the classical optimizers can be split into two categories. The first one is the gradient-based optimizer, which navigates the landscape iteratively using the direction indicated by the gradient as a compass. As the observables of VQAs are estimated by statistical estimators, they fit under the general method of Stochastic Gradient Descent (SGD). By the natural similarity of VQAs with machine learning, some used SGD on VQAs are inspired or imported from machine learning, such as Adam [90] and the individual Coupled Adaptive Number of Shots (iCANS) [91]. On the other hand, there are approaches that do not use gradient (at least directly) called gradient-free [92–94].

The aforementioned barren plateau is a phenomenon that occurs in many classes of VQAs in which the cost landscape becomes flatter as we increase the number of qubits of the quantum circuit, meaning that the gradient becomes exponentially small and the optimization process becomes inert. Following the description of Blekos et al. [10], formally, the barren plateau phenomena arise for a cost function $c(\boldsymbol{\theta})$ if for all optimization parameters $\theta_j \in \boldsymbol{\theta}$,

$$\mathrm{Var}\left[\frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j}\right] = \mathcal{O}(b^{-n}), \tag{3.10}$$

where n is the number of qubits and b a constant such as $b > 1$. Since the variance is a concept—as we discussed in Subsec. 2.4.2—related to the spread from the mean, the gradient of the cost function is exponentially small on average. Furthermore, Chebyshev's inequality finds importance in this context in such a way that its application gives

$$\mathrm{P}\left[\left|\frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j}\right| \geq k\right] \leq \frac{1}{k^2}\mathrm{Var}\left[\frac{\partial c(\boldsymbol{\theta})}{\partial \theta_j}\right] = \mathcal{O}(b^{-n}) \tag{3.11}$$

for a given $k > 0$. The implication of the exponential decay of the probability of the absolute value of gradient being equal to or greater than k is a flatter landscape.

The barren plateau phenomena affect not only gradient-based but also the gradient-free approaches [95]. There is a connection between BPs and the randomness of the ansatz in such a way that if an ansatz forms a 2-design, i.e., matches the uniform distribution of unitaries up to the second statistical moment, it presents BPs [89]. In the presence of noise, the problem is more relentless since barren plateaus appear independently of the ansatz used [96].

The barren plateau problem draws attention to the issue of parameter initialization. The most common strategy, random initializing, can lead to an unfavorable region, such as regions of barren plateaus. However, in some situations, such as in Zhou et al. [97] paper, heuristic strategies based on empirical observation of the optimal parameters can be used to get better results than random initialization.

3.4.3 Applications

A notable advantage of the VQAs framework is its applicability, widely discussed in Cerezo et al. [4] paper. The range of tasks that VQAs can tackle is quite wide, much broader than the context of combinatorial optimization. Indeed, VQAs even support universal quantum computing [98]. The two main classes of VQAs are the Quantum Approximate Optimization Algorithm, applied mainly on combinatorial optimization, discussed in Sec. 3.6, and the Variational Quantum Eigensolver (VQE), originally introduced by Peruzzo et al. [99] and subsequently improved and extended by McClean et al. [100]. The VQE is a most direct application of the variational principle, used to find the ground state of quantum systems in the physics and chemistry context.

Among the VQAs for specific tasks, we cite for mathematical applications the Variational Quantum Factoring (VQF) [101] and the Variational Quantum Linear Solver (VQLS) [102], NISQ alternatives for the Shor's algorithm [103, 104] on integer factorization problem and the HHL algorithm [79] on the problem of linear system of equations, respectively. Other general applications of VQAs are compilation of quantum circuits, dynamical simulations, error correction, machine learning, and quantum information.

3.5 Quantum Adiabatic Algorithm

The Quantum Approximate Optimization Algorithm is derived from the Quantum Adiabatic Algorithm (QAA), which, in turn, fits under the more general universal model of quantum computing called Adiabatic Quantum Computation (AQC) [105]. The idea behind QAA came from a fundamental result of quantum mechanics, the adiabatic theorem, originally stated by Born and Fock [106]. Consider that we are interested in finding the ground state of a Hamiltonian H_C, of difficult preparation. In a combinatorial optimization context, H_C codifies the solution space of a given COP. The QAA, introduced on the papers [7, 8], consists in preparing the system on the known ground state of a second Hamiltonian H_M, of easy preparation and that does not commute with H_C, and then evolving continuously on the time τ [6] the system from H_M to H_C on a transitional time-dependent Hamiltonian $H(\tau)$ by the interpolation

$$H(\tau) = f(\tau)H_C + g(\tau)H_M \tag{3.12}$$

for $\tau = [0, \tau_{max}]$, where $f(\tau) = \tau/\tau_{max}$ and $g(\tau) = 1 - \tau/\tau_{max}$. By the adiabatic theorem, if the evolution is slow enough—that is, sufficiently large τ_{max}—the system keeps on its ground state throughout the process, and we find the desired ground state of H_C at the end of interpolation.

A condition for the adiabatic theorem to be applicable is that there must be an energy gap between the ground state and the first excited state. The smaller the gap, the slower the evolution must be to avoid "mixing" between both states, since the quantum system needs a certain amount of time to "adapt" to external disturbances. Furthermore, as emphasized in the original article of QAA [7], H_C and H_M must not commute. Otherwise, from a well-known result of linear algebra, they share the same eigenvectors, changing only the eigenvalues. That way, at some point in the interpolation, the energy gap would become zero and therefore the algorithm would fail.

Following the description Blekos et al. [10], the evolution unitary is defined as $U(\tau) = e^{-i\int_0^\tau H(u)du}$. Since H_C and H_M do not commute, we compile approximately the continuous evolution of the evolution unitary on the gate model quantum computing by the Trotterization [107] process. That way, from the Trotter-Suzuki formula on r steps,

$$\begin{aligned} U(\tau) &\approx \prod_{k=0}^{r-1} \exp\left[-iH(k\Delta\tau)\Delta\tau\right] \\ &= \prod_{k=0}^{r-1} \exp\left[-if(k\Delta\tau)H_C\Delta\tau\right]\exp\left[-ig(k\Delta\tau)H_M\Delta\tau\right], \end{aligned} \tag{3.13}$$

where $\Delta\tau = \tau/r$.

[6] We do not use the usual symbol t for the time to avoid confusion with the threshold value of GM-Th-QAOA, introduced on of Subsec. 3.8.2.

3.6 Quantum Approximate Optimization Algorithm

The Quantum Approximate Optimization Algorithm (QAOA) was originally introduced by Farhi, Goldstone, and Gutmann [5] as a heuristic VQA to Max-Cut problem in the NISQ context, being until today likely the most well-studied problem in this context [10]. However, that original framework is directly applicable to unconstrained optimization problems and indirectly applicable to constrained optimization by adding “penalties” to the cost function on states considered unfeasible, compromising the performance of the algorithm—see Slate et al. [25] for an example of application on constrained optimization context. QAOA is a modification of QAA on the Trotterization form of Eq. (3.13), replacing the functions $f(\tau)$ and $g(\tau)$ by variational parameters to be optimized. Specifically, for all k such that $1 \leq k \leq r$, $f((k-1)\Delta\tau)\Delta\tau$ becomes the parameter γ_k and $g((k-1)\Delta\tau)\Delta\tau$ becomes the parameter β_k. Thus, we introduce $2r$ parameters in the vectors $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$.

In particular, the Hamiltonian H_C, diagonal in the computational basis, encodes Max-Cut or other unconstrained problems on n qubits such that

$$H_C|k\rangle = c(k)|k\rangle, \tag{3.14}$$

where $c(k)$ is the cost function and $|k\rangle$ is the quantum state encoding the solution given by a n-bit string $\boldsymbol{k}$. The spectrum of the Hamiltonian H_C corresponds to the solution space of the problem. The Hamiltonian H_M, in turn, is given by

$$H_M = -\sum_{j=1}^{n} \sigma_j^X, \tag{3.15}$$

where σ_j^X is the Pauli-X operator applied on the jth qubit. The ground state of H_M is given by the uniform superposition $|+\rangle^{\otimes n} = \frac{1}{\sqrt{2^n}} \sum_{k=0}^{2^n-1} |k\rangle$, where $\otimes$ denotes the Kronecker product. Thus, the evolution of QAOA with r layers, given by the state $|\psi^{(r)}\rangle$, is the alternate application the operators $U_P(\gamma) = e^{-i\gamma H_C}$ and $U_M(\beta) = e^{-i\beta H_M}$ on the initial state $|+\rangle^{\otimes n}$, that is,

$$|\psi^{(r)}\rangle = U_M(\beta_r)U_P(\gamma_r)\ldots U_M(\beta_1)U_P(\gamma_1)|+\rangle^{\otimes n}. \tag{3.16}$$

We optimize the parameters in the variational manner aiming to minimize[7] the

[7] The original definition of Farhi, Goldstone, and Gutmann [5] is for maximization problems, in which H_M is given without the minus sign on Eq. (3.15).

expectation value $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$, given by

$$\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle = \sum_{k=0}^{2^n-1} |\alpha_k|^2 c(k), \tag{3.17}$$

where α_k is the amplitude of the state $|k\rangle$. The search space of β_j can be restricted to $(0,\pi]$ and, if the cost function has only integer values, such as the Max-Cut problem case, γ_j can be restricted to $(0,2\pi]$.

The expectation value can be statistically estimated with an estimator from the solutions obtained by the measurement of the state $|\psi^{(r)}\rangle$. Marsh and Wang [14] show that if the optimization problem is an NP optimization problem polynomially bounded—that is, an NP optimization problem with the values of the cost function bounded by a polynomial function on the size of the instance—we can obtain the expectation value on a fixed confidence interval with a random sample of polynomial size. Furthermore, follows that the minimum expectation value of r layers is smaller or equal to the minimum expectation value of $r-1$ layers and that the minimum expectation value on $r \to \infty$ gives the optimal solution.

The compilation of the quantum circuit of $U_M(\beta)$, in addition to the $U_P(\gamma)$ for Max-Cut, can be found in details on Blekos et al. [10] review paper.

3.7 Quantum Alternating Operator Ansatz

The Quantum Alternating Operator Ansatz (QAOA), introduced by Hadfield et al. [6], is a generalization of the Quantum Approximate Optimization Algorithm. While the last is directly applicable only to unconstrained optimization problems, the first generalizes the framework to a constrained optimization context. Instead of acting as a Hilbert space of dimension 2^n, where n is the number of qubits, the Quantum Alternating Operator Ansatz acts in a generic subspace of M feasible solutions of the problem, where M is not necessarily a power of 2. The operators are also generalized: $U_M(\beta)$ beyond the Hamiltonian of Eq. (3.15) and $U_P(\gamma)$ beyond directly codifying the cost function.

As emphasized by Hadfield et al. [6], the main application of QAOA is to heuristic[8] optimization of NP-Hard optimization problems. However, also can be used, for instance, to exact optimization [74, 75, 108]. As aforementioned—and indicated in its terminology—the QAOA can also be seen as an ansatz in the context of VQAs, being applicable, for instance, to VQAs of specific purpose such as VQF [101] and VQLS [102]. Focusing on combinatorial optimization context, we define QAOA as follows.

[8] Approximation ratio guarantees are rare in the context of QAOA, as discussed on Subsec. 3.7.2.

Definition 17 (Quantum Alternating Operator Ansatz) *Consider an instance of a combinatorial optimization problem defined on a domain S with a cost function (the objective function) $c(k) : S \to \mathbb{R}$ be minimized. For some Hilbert space known as configuration space, the algorithm acts in some subspace called feasible subspace spanned by $M = |S|$ basis states (the feasible states) codifying the solutions of S. The state final of QAOA, denoted $|\psi^{(r)}\rangle$, is given by*

$$|\psi^{(r)}\rangle = U_M(\beta_r)U_P(\gamma_r)\ldots U_M(\beta_1)U_P(\gamma_1)|\psi\rangle. \tag{3.18}$$

Here,

- *r is the number of rounds/layers/iterations*[9] *or the depth of QAOA;*
- *$|\psi\rangle$ is a generic initial state;*
- *$U_P(\gamma) = e^{-i\gamma H_Q}$ is the phase separation operator (or phase separator), where H_Q is a Hamiltonian that encodes a real-valued function $q(k)$ compiled from the cost function such that*

  $$H_Q|k\rangle = q(k)|k\rangle \tag{3.19}$$

  for any feasible state $|k\rangle$;
- *$U_M(\beta) = e^{-i\beta H_M}$ is the mixing operator (or mixer operator), where H_M is mixer Hamiltonian (or mixing Hamiltonian or even driver Hamiltonian);*
- *sets $\boldsymbol{b} = (\beta_1, \ldots, \beta_r)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_r)$ are the optimization parameters (or angles), with search space depending on the mixing operator and phase separation operator, respectively.*

To this work, we assume that the goal of QAOA is to minimize the expectation value $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$. The diagonal Hamiltonian H_C called the Hamiltonian problem is defined analogously as on Quantum Approximate Optimization Algorithm with $H_C|k\rangle = c(k)|k\rangle$ for any feasible state $|k\rangle$. Furthermore, we also assume H_Q is diagonal on the computational basis and that the Hilbert space of the configuration space is defined on a qubit system.

3.7.1 Phase separation and mixing operators

The mixing operator depends on the structure of the combinatorial, while the phase separation operator depends on the cost function. For a given combinatorial domain

[9] For this work, these terminologies always refer to r and not to the variational loop of the classical optimizer of QAOA.

corresponding to a given feasible subspace, there are many possibilities of mixing operators, while for a given cost function, there are a variety of possible compilations for the phase separation operator. Because of that, we say that $U_M(\beta)$ and $U_P(\gamma)$ are families of operators.

Beyond the original interpretation of Quantum Approximate Optimization Algorithm from the points of view of the QAA, we can interpret the operators of Quantum Alternating Operator Ansatz as follows [11, 12]. The phase separation operator changes the relative phases between states by introducing bias according to the cost function to be optimized, acting as the "oracle" of the problem. The mixing operator is responsible for generating interference between the states, "mixing" its amplitudes to amplify the probability of measuring states corresponding to high-quality solutions, i.e., desirable solutions from a heuristic algorithm point of view.

In the original Quantum Approximate Optimization Algorithm framework, the mixer Hamiltonian is the mixer of Eq. (3.15) (without the minus sign), called transverse field mixer (or single-qubit-X mixer); the phase separation encodes precisely the cost function, that is, $q(k) = c(k)$ and $H_Q = H_C$; and the initial is given by $|+\rangle^{\otimes n}$. In that case, the configuration space is equal to the feasible subspace, which, in general, is straightforward on unconstrained optimization. The transverse field mixer connects pairs of states with unit Hamming distance, that is, $\langle x|H_C|y\rangle = 1$ if the Hamming distance of the binary representations of x and y is equal to 1 and $\langle x|H_C|y\rangle = 0$ otherwise.

The phase separation operator, the problem Hamiltonian, and the Hamiltonian H_Q can be written using projectors as

$$U_P(\gamma) = \sum_{k\in S} e^{-i\gamma q(k)}|k\rangle\langle k|,\ H_C = \sum_{k\in S} c(k)|k\rangle\langle k|,\ H_Q = \sum_{k\in S} q(k)|k\rangle\langle k|. \tag{3.20}$$

Provided that the function $q(k)$ is efficiently computable, we can compile efficiently the phase separation operator $U_P(\gamma)$ with the procedure of Childs [109] (see the quantum circuit of Figure 1-1 on the thesis of Childs [109]).

Although there are exceptions such as GM-Th-QAOA, discussed on Subsec. 3.8.2, in almost all cases, $q(k) = c(k)$. On the other hand, concerning the mixing operators, Hadfield et al. [6] define several families that apply to a wide range of optimization problems. We highlight the family of XY-mixers, which are a sum of the 2-local operators $\sigma_j^X\sigma_k^X+\sigma_j^Y\sigma_k^Y$[10], where j and k are arbitrary qubits. Each operator behaves exactly as a SWAP gate on the subspace spanned by $\{|0_j1_k\rangle, |1_j0_k\rangle\}$, preserving the Hamming weight. That naturally induces application on problems defined on the combinatorial domain of the Hamming weight k bit strings, such as Max k-Vertex

[10] The operator σ_j^Y is the Pauli-Y gate applied to the jth qubit.

Cover, k-Densest Subgraph, Max Bisection [11, 28]. Two prominent mixers of the XY-mixers family are the ring and clique (complete graph) mixers, given for n qubits by

$$H_M = \sum_{j,k:k=j+1 \mod n} \sigma_j^X \sigma_k^X + \sigma_j^Y \sigma_k^Y, \; H_M = \sum_{j,k:k>j} \sigma_j^X \sigma_k^X + \sigma_j^Y \sigma_k^Y, \tag{3.21}$$

respectively. Ring mixer sum over cyclically adjacent qubits and clique mixer sum over all pairs of qubits. Other mixers were introduced later, such as line mixer [33], given by

$$H_M = \sum_{j,k:|k-j|=1} |j\rangle\langle k| + |k\rangle\langle j|, \tag{3.22}$$

which connects states with a metric based on the arithmetic distance of feasible solutions, and the Grover mixer, discussed in the next section.

3.7.2 Angles finding and analytical results

The optimal (or at least near-optimal) angles of QAOA are obtained in the majority of cases with the usual but costly outer loop by using classical optimizers, such as Nelder-Mead [110] and Broyden Fletcher, Goldfarb, Shanno (BFGS) [111]. However, in some very particular cases, such as in this work for Grover mixer variants, it is possible to compute the parameters analytically [74, 112–114].

In fact, as discussed by Golden et al. [11], the difficulty of analytically obtaining the optimal parameters is a consequence of the more general issue that analytical results are historically rare and sparse in QAOA literature due to the high complexity of quantum operators. Furthermore, the number of optimization parameters grows with the number of layers, making it difficult to generalize methods beyond a small number of layers. Thus, little is known about the theoretical performance[11] of QAOA and its potential when compared to the classical algorithms, being much of the knowledge based on numerical evidence, also limited by the inherent challenges of the classical simulation of quantum circuits. As Golden et al. [11] mention, the most well-known analytical result of QAOA is the approximation ratio guarantees on 3-regular graphs of the Max-Cut problem for a small number of layers [5, 115, 116]. In general, approximation ratio guarantee is likely the most common performance metric in the analytical results of QAOA literature, a topic discussed in detail in Blekos et al. [10] review paper. In the present dissertation, we provide bounds on the performance of QAOA on Grover mixer variants with the alternative metrics of the standard score and the quantile of the solution space.

Other aspects that can be considered in QAOA research, such as computational

[11] The performance of QAOA throughout this work refers to the quality of the result obtained by the algorithm provided from some metric, such as the approximation ratio.

resource efficiency, noise and error considerations, and hardware-specific approaches, all covered by Blekos et al. [10] review, are outside the scope of this work. Throughout this work, we consider QAOA in its ideal conditions, without considering the noise or the complexity of the compilation on the algorithm in a quantum circuit.

3.8 Grover mixer

The Grover mixer Hamiltonian is given by

$$H_M = |s\rangle\langle s|, \tag{3.23}$$

where $|s\rangle$ is a uniform superposition over all feasible states, that is,

$$|s\rangle = \frac{1}{\sqrt{M}} \sum_{k \in S} |k\rangle. \tag{3.24}$$

It can be shown that the mixing operator of Grover mixer Hamiltonian, called Grover mixer operator or simply Grover mixer, is given by

$$U_M(\beta) = \mathbb{I} + B(\beta)|s\rangle\langle s|, \tag{3.25}$$

where $B(\beta) = -1 + e^{-i\beta}$. Taking $\beta = \pi$ reduces the Grover mixer to Grover's diffusion operator of Eq. (3.4) up to a global phase, justifying its terminology. The period of this operator is 2π. In particular, here we set the search space of β as $(-\pi, \pi]$. The search space of γ depends on the optimization problem. In the general case, $\gamma \in \mathbb{R}$, while assuming integer costs, we can restrict it to $(-\pi, \pi]$.

Unlike the traverse field mixer, which connects the states of unit Hamming distance, the Grover mixer connects all pairs of states. This characteristic is central to this work and culminates in the fact that the Grover mixer is invariant over any permutation of states, discussed in Chapter 4. Another significant property of the Grover mixer is that assuming the initial state is $|s\rangle$, degenerate solutions—solutions with the same cost—share the same amplitudes throughout the algorithm [13].

3.8.1 Grover Mixer Quantum Alternating Operator Ansatz

The Grover Mixer Quantum Alternating Operator Ansatz (GM-QAOA) was introduced formally by Bärtschi and Eidenbenz [13]. The Grover mixer had already been used at least by Morales, Tlyachev, and Biamonte [75], Akshay et al. [16], and Sundar et al. [117]. All mentioned situations are in an unconstrained context. The framework of Bärtschi and Eidenbenz [13] generalizes Grover mixer to include constrained optimization. Here, we define GM-QAOA as follows.

Definition 18 (GM-QAOA) *GM-QAOA is the particular case of QAOA on Def. 17 in which the mixer Hamiltonian is the Grover mixer, given by Eq.* (3.23)*; the phase separation operator codifies the own cost function, that is,* $H_Q = H_C$*; the initial state is* $|s\rangle$*, given by Eq.* (3.24)*; and the goal is to minimize expectation value* $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$*, particularly denoted* $E_r(\boldsymbol{\beta},\boldsymbol{\gamma})$*.*

The framework of Bärtschi and Eidenbenz [13] gives an efficient algorithm to any NP optimization problem which admits an efficient preparation of the uniform superposition $|s\rangle$, which encompasses several optimization problems. Beyond the trivial case of unconstrained problems in which $|s\rangle = |+\rangle^{\otimes n}$, we can cite the problems with combinatorial domain on Hamming weight k bit strings (the optimization problems in which the XY-mixers can be used), the problems with combinatorial domain on permutations such as Traveling Salesman Problem, and the Discrete Portfolio Rebalancing.

3.8.2 Grover Mixer Threshold Quantum Alternating Operator Ansatz

Before defining the Grover Mixer Threshold Quantum Alternating Operator Ansatz (GM-Th-QAOA), we must consider the Threshold QAOA (Th-QAOA). Both Th-QAOA and GM-Th-QAOA are introduced by Golden et al. [26]. Adapting the original definition to consider minimization problems, the Th-QAOA is a variant of QAOA in which the phase separation operator, instead codifies, as usual, the cost function, it codifies the compilation of the $c(k)$ into the threshold function given by

$$T_h(k) = \begin{cases} -1, & c(k) \leq t \\ 0, & \text{otherwise}, \end{cases} \tag{3.26}$$

for a threshold value t that must be optimized. We can consider as candidates of optimal threshold all possible values of the cost function except the maximum cost, which results in a trivial compilation. Combining Th-QAOA with the choice of Grover mixer as the mixer Hamiltonian gives the GM-Th-QAOA, which leads to the following definition.

Definition 19 (GM-Th-QAOA) *GM-Th-QAOA is the particular case of QAOA on Def. 17 in which the mixer Hamiltonian is the Grover mixer, given by Eq.* (3.23)*; the phase separation operator codifies the function* $q(k) = T_h(k)$*; the initial state is* $|s\rangle$*, given by Eq.* (3.24)*; and the goal is to minimize expectation value* $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$*, particularly denoted* $E_r(t)$*.*

One significant aspect of the GM-Th-QAOA is that for a fixed threshold t, it can emulate the execution of Grover's algorithm with the marked elements being states k such that $c(k) \leq t$. To notice that, recall that $\beta = \pi$ reduces Grover mixer to Grover's diffusion operator up a global phase. In GM-Th-QAOA, additionally, taking $\gamma = \pi$ reduces the phase separation to Grover's oracle of Eq. (3.3) for the claimed marked elements. Combining it with the initial condition of uniform superposition over all states, if we set π for both angles on all the layers, we emulate Grover's algorithm. Furthermore, note that the procedure of applying Grover's algorithm to a solution space split by a threshold function on GM-Th-QAOA resembles GAS. However, both algorithms are conceptually distinct since the metric of the choice of threshold on GM-Th-QAOA is the expectation value, while the threshold used on GAS is the best value known so far, obtained iteratively.

Optimal angles of GM-Th-QAOA

Since degenerate solutions share the same amplitude on Grover mixer variants, the final state of GM-Th-QAOA for r rounds can be written as

$$|\psi^{(r)}\rangle = c_1^{(r)} \sum_{k \in S : c(k) \leq t} |k\rangle + c_0^{(r)} \sum_{k \in S : c(k) > t} |k\rangle, \tag{3.27}$$

where $c_1^{(r)}$ and $c_0^{(r)}$ are generic amplitudes for states below/equal and above the threshold value, respectively. We denote by ρ the ratio of states below/equal on the entire domain of S. For a given t, ρ is fixed and we minimize the expectation value of GM-Th-QAOA by maximizing $|c_1^{(r)}|^2$—or equivalently minimizing $|c_0^{(r)}|^2$.

Golden et al. [26] find analytically the optimal angles β and γ [12] for $r = 1$. Adapting to our notation and minimization problems, we have $\beta = \gamma = \pi$ if $\rho \leq 0.25$—reducing the operators to a Grover iteration—and

$$\beta = -\gamma = \arctan\left(-\sqrt{4\rho - 1}, 2\rho - 1\right) \tag{3.28}$$

otherwise, where the function $\arctan(a, b)$ for $a, b \in \mathbb{R}$ calculates arc tangent considering the quadrant. In particular, for $r > 0.25$, the optimal angles maximize the amplitude of $|c_1^{(1)}|^2$ as maximum as possible since it gives $|c_0^{(r)}|^2 = 0$. For arbitrary r, Golden et al. [26] conclude that set the angles $\beta_j = \gamma_j = \pi$ for all $j < r$ and

$$\begin{aligned} \beta_r &= \arctan\left(-\sqrt{\Delta}|c_0^{(r-1)}|, \frac{2\rho}{M} - (c_0^{(r-1)})^2\right), \\ \gamma_r &= -\arctan\left(-\frac{\sqrt{\Delta}}{c_1^{(r-1)}\,\mathrm{sgn}\,(c_0^{(r-1)})}, \frac{c_{0,\pi}^{(r-1)}(2r-1)}{c_1^{(r-1)}}\right), \end{aligned} \tag{3.29}$$

[12] For QAOA with a single layer, we simplify the notation with $\beta = \beta_1$ and $\gamma = \gamma_1$.

[13] where

$$\Delta = \frac{4\rho}{M} - (c_0^{(r-1)})^2, \tag{3.30}$$

is a optimal choice of angles for GM-Th-QAOA if $\Delta > 0$. In that case, follows $|c_0^{(r)}|^2 = 0$. Golden et al. [26] numerically observe that set $\beta_j = \gamma_j = \pi$ for all j gives the minimum expectation value if the condition $\Delta > 0$ is not satisfied. We prove in Subsec. 4.1.4 that these angles are indeed optimal.

Efficient parameter finding

An advantage of GM-Th-QAOA over GM-QAOA is that it admits an efficient method for finding the parameters (the angles $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the threshold t) that eliminates the costly outer loop parameter finding of QAOA. The method, introduced by Golden et al. [26], is based on the previous results, and for a fixed r, has complexity given by $\mathcal{O}(\log(r)\log(t_{\text{dif}}))$, where t_{dif} is the number of non-degenerate costs of the cost function. The factor $\log(r)$ concern about finding the angles $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, while $\log(t_{\text{dif}})$ concern about finding the threshold.

The optimization procedure of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, for a fixed threshold t, consists in find out if there a transition round τ such that $\tau \leq r$ in which $\beta_j = \gamma_j = \pi$ for all $j < \tau$ and the angles of Eq. (3.29) on the τth layer gives the optimal $|c_0^{(\tau)}|^2 = 0$. If so, we set, in addition to that angles, $\beta_j = \gamma_j = 0$ for all $j > \tau$ to make the operators trivial. Otherwise, we set $\beta_j = \gamma_j = \pi$ for all j. The optimality of the angles of the last case, hitherto based on numerical evidence, is proved in the present work. The procedure, which resembles the exponential quantum search, is the following.

- Make a exponential search over the number of rounds $r_k = \lceil K^k \rceil$ for $K > 1$ using the angles $\beta_j = \gamma_j = \pi$ for all j to find values such that r_{k+1} rounds gives a lower expectation than r_{k+2} rounds. These values exist if the transition round exists;

- Make a binary search on the interval r_k between r_{k+2} to find the round r_o which gives the minimum expectation value with angles $\beta_j = \gamma_j = \pi$ for all j;

- Since r_0 could overshoot or undershot the optimal number of layers, we must test either $\tau = r_o$ or $\tau = r_o + 1$ if the transition round.

In particular, if the number of layers is small, it would be more efficient to do a linear search instead of the previous procedure to find the transition layer.

The method of finding the best threshold t between all t_{dif} candidates is based on the numerical observation that the curve of the expectation value versus the

[13] The real-valued function $\operatorname{sgn}(x)$ is the sign function, such that $\operatorname{sgn}(x) = -1$ if $x < 0$, $\operatorname{sgn}(x) = 0$ if $x = 0$, and $\operatorname{sgn}(x) = 1$ otherwise.

threshold value for angles $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ obtained on the previous procedure decreases monotonically up to a valley value and then increases monotonically—recall that we are considering minimization problems. Observe, for instance, the Fig. 1 of Golden et al. [26] paper. That way, conjecturing that behavior holds in general, we can apply an analog procedure of the one used to find the transition round of the angles finding, but now, with the "transition" being the threshold value of the minimum expectation value. If that does not hold generally, would be required a linear search to find the threshold, which represents an exponential loss. Although usually t_{dif} is bounded on combinatorial optimization problems, such as the numbers of edges plus one in Max-Cut, an exponential gain is always desirable. We prove that conjecture in Subsec. 4.2.1.

3.8.3 Classical simulation of QAOA with Groxer mixer

In addition to introducing Th-QAOA and GM-Th-QAOA, the Golden et al. [26] paper introduce a method of classical simulation of QAOA with Grover mixer—which includes both GM-QAOA and GM-Th-QAOA—that allows simulate the GM-Th-QAOA with $16,384$ rounds on instances that would require 100 qubits, far beyond the limit classical simulation of quantum circuit in the general case. In the GM-QAOA case, which, on the other hand, does not have an efficient parameters finding, the simulation goes to 20 rounds and 40 qubits. Although the scale is much smaller than GM-Th-QAOA, it is beyond current general QAOA simulations— until the knowledge of the author. In that simulation method, instead of decomposing the unitaries into quantum gates, we algebraically calculate the expectation value. Here, we provide an intuitive notion of the concept on which its the method is based. Precise details can be found in the original paper.

The method takes advantage of the fact that degenerate solutions share the same amplitudes, grouping these solutions and thus working in the subspace of non-degenerate solutions, which has dimension t_{dif} (recall that t_{dif} denoted the number of non-degenerate costs of the cost function). To get that, we make a preprocessing in which we compute via brute force all feasible solutions and then build a reduced space with two sets, one that stores all non-degenerate solutions and the other with the numbers of solutions with each cost. In that subspace, we can simulate the matrix dynamics of the application of the unitaries $U_P(\gamma)$ and $U_M(\beta)$ on an arbitrary state with complexity linear on t_{dif} for both time and space. For $U_P(\gamma)$, since that operator is diagonal, it is trivial to compute its dynamics linearly. On the other hand, for $U_M(\beta)$, we present the intuition of the simpler case of the original feasible space. For a generic state $|\alpha\rangle$,

$$U_M(\beta)|\alpha\rangle = |\alpha\rangle + B(\beta)\langle s|\alpha\rangle|s\rangle. \tag{3.31}$$

Eq. (3.31) can be computed with linear time and space on the dimension of the feasible subspace M since the inner product $\langle s|\alpha\rangle$ has linear complexity on M. On the reduced subspace, we can simulate it by taking a sum ponderated by the "weight" of each cost, that is, the number of degenerate solutions stored on the second mentioned set obtained in the preprocessing—a procedure that also has linear complexity on the space dimension.

The number of non-degenerate solutions t_{dif} can be exponentially smaller than the number of feasible solutions M. For instance, in Max-Cut while $M = 2^{|\mathcal{V}|}$, t_{dif} is in the worst case $\mathcal{O}(|\mathcal{V}|^2)$, since it is bounded by the numbers of edges. In constrained problems, the gain from the circuit simulation can be even greater since the configuration space is larger than the feasible subspace. In these cases, the bottleneck of the simulation is the number of solutions we can compute on the preprocessing, allowing the simulation of a huge number of layers since the non-degenerate subspace is small.

3.9 Quantum Walk-based Optimization Algorithm

The Quantum Walk-based Optimization Algorithm (QWOA) is a generalization of the Quantum Approximate Optimization Algorithm that introduces a new interpretation for the algorithm. The mixing operator $U_M(\beta) = e^{-i\beta H_M}$, from QWOA view, is a Continuous-time Quantum Walk (CTWQ) [20–22] operator of time β on the graph with Laplacian or adjacency matrix given by H_M. A CTWQ is the quantum analog to the classical continuous-time Markov chain [118]. Details about CTQWs and general quantum walks can be found in Portugal's [22] book. The transverse field, the original mixer, is a CTWQ on a hypercube graph.

The original formulation of QWOA is from Marsh and Wang [14]. In it, the CTQW is on a modified hypercube graph, that is, a subgraph of the hypercube graph with vertices corresponds to the feasible solutions. That framework includes efficient circuits for any polynomial bounded NP optimization problem, which is particularly interesting for problems with combinatorial domains naturally difficult to connect with mixing operators, such as the Minimum Vertex Cover. Subsequently, Marsh and Wang [12] introduce a new formulation for QWOA, considering now CTQW on the class of circulant graphs [119], with adjacency matrix diagonalizable by the efficient Quantum Fourier Transform (QFT). That formulation requires an efficient indexing function, i.e., a bijection $id : S \to \{0, 1, \ldots, M-1\}$ that associate uniquely each element of S with a numerical index[14]. Furthermore, the inverse $id^{-1} : \{0, 1, \ldots, M-1\} \to S$, called un-indexing function, also must be efficient. Analog quantum operators can be built with these functions, and the CTQW can be

[14] Also known in the literature as ranking.

implemented in an indexed subspace. The paper of Marsh and Wang [12] brings efficient indexing functions for permutations, states with k-combinations, and lattice paths. Indexing functions for other several families of combinatorial objects can be found in Loehr's [120] book. Later works considered numerical studies of QWOA on the problems of Capacitated Vehicle Routing [24] and the Portfolio Optimization [25].

A graph with particular interest is the complete graph, which has Laplacian matrix $\mathcal{L} = M(\mathbb{I} - |s\rangle\langle s|)$ and mixing operator, up to a global phase,

$$U_M(\tau) = \mathbb{I} + (-1 + e^{iM\tau})|s\rangle\langle s|. \tag{3.32}$$

A quantum circuit to implement the operator of Eq. (3.32) that is distinct from the framework of circulant graphs can be found in Bennett et al. [24]. Taking $\beta = -\tau M$ gives exact the mixing operator of Eq. (3.25). As M is constant for a given instance, QWOA on the complete graph and GM-QAOA are equivalent, up to a scale change on the parameter—that equivalence was noted by Bennett and Wang [23]. Thus, the framework of QWOA on the complete graph can be viewed as an alternative implementation of the Grover mixer operator. Furthermore, the intuition provided by the complete graph can be particularly useful to view the properties of the Grover mixer, as emphasized in Chapter 4.

3.10 Maximum Amplification Optimization Algorithm

The Maximum Amplification Optimization Algorithm (MAOA) is an algorithm introduced by Bennett and Wang [23] to correct a weakness of QWOA on the complete graph. Specifically, it was observed in the previous work of Bennett et al. [24] that the algorithm tends to amplify the probability of near-optimal solutions more than the probability of optimal ones. To bypass this issue, Bennett and Wang [23] conclude that the amplification of the probability of the optimal solution can be a more effective metric than the usual expectation value of QAOA. Then, systematic numerical experiments were done to determine the combination of the graph of the CTQW and the number of non-degenerate solutions codified on the phase separation in which the probability of measuring a given state is most amplified. The conclusion is that the best combination is the complete graph with 2 degenerate solutions—a function similar to $T_h(k)$—and choices of angles equal to π. As discussed for GM-Th-QAOA, that combined choice reduces the algorithm to Grover's algorithm.

From these findings, MAOA is developing to act approximately on the low-convergence regime of Grover's search. That is because, recall from Subsec. 3.2.2, that the largest amplification of Grover's algorithm is provided on a low-convergence regime with $\eta = (2r + 1)^2$. The algorithm, whose details are omitted here, consists of two steps. Firstly, we estimate efficiently, for a given number of layers r, the threshold value in which the final probability of r Grover's iterations is close to 1/40 (to get accurate within 1% with respect to the low-convergence regime). Then, we make repeated measurements of r Grover's iterations with the obtained threshold to get high-quality solutions with a rate of $\approx 1/40$.

The performance of MAOA was compared directly with a modified version of the GAS so-called restricted Grover Adaptive Search (RGAS), in which the procedure is the same as GAS, except that the maximum allowed number of iterations of Grover's algorithm is restricted. Considering the same number of Grover's iterations, both algorithms are compared in terms of the probability of measuring an optimal solution in the function of the computational effort for the problems of Capacitated Vehicle Routing and Portfolio Optimization, and for arbitrarily large instances with solution space normally distributed[15]. MAOA consistently overcomes RGAS, although reflecting the same quadratic Grover-like speed-up.

It is worth mentioning that Bennett and Wang [23] paper introduces a method of analysis of QWOA on the complete graph with phase separation codifying a function with 2 non-degenerate solutions. Specifically, that method uses the degeneracy on the solution space to make edge contractions on the complete graph of QWOA and work on a reduced subspace of 2 dimensions. In a certain sense, the statistic approach of Chapter 4 is a generalization of this analysis method on a solution space with an arbitrary number of non-degenerate solutions codified by its probability distribution.

3.11 Performance of QAOA with Grover mixer

The individual and comparative performance of QAOA with Grover mixer, which is the main motivation of this work, has been considered in the literature on some occasions. Firstly, within its variants, in numerical experiments, the performance of GM-Th-QAOA consistently overcomes GM-QAOA in all instances considered. Beginning on the original paper of GM-Th-QAOA [26] on the problems of Max-Cut, Max k-Vertex Cover, k-Densest Subgraph, and Max Bisection, and later on subsequent works with Max 2-SAT, Max 3-SAT problems [17] and Max k-Vertex Cover, k-Densest Subgraph, and Max Bisection problems [11]. It is an open question if the superior performance of GM-Th-QAOA over GM-QAOA holds in general.

[15] In Chapter 4 we provide a precise definition for the probability distribution of the solution space of combinatorial optimization problems.

Concerning comparative performance, especially with the Grover transverse, the initial thought was that the Grover mixer would have better performance since, as it admits all state transitions, the mixing process is faster, and the symmetry among states is global [12, 16, 25]. The numerical experiments given by Akshay et al. [16] on the unconstrained problems Max 2-SAT and Max 3-SAT with 6 qubits corroborated that argument with the Grover mixer (GM-QAOA) having a superior performance than the transverse field mixer. However, the latter experiment of Golden et al. [17] went against the initial thought. In particular, the numerical study of Akshay et al. [16] was scaled by Golden et al. [17] up to 14 qubits, and the situation has reversed, with both GM-QAOA and GM-Th-QAOA doing worse than the transverse field mixer on both Max 2-SAT and Max 3-SAT problems. Although, of course, that does not imply that the results can be automatically generalized to even larger instances, there is a theoretical argument in this direction: by its global symmetry among the vertices, Grover mixer does not see the structure of the problem[16], being possibly limited to the bound of the unstructured search problem. In other words, QAOA with Grover mixer would be limited by a quadratic speed-up over classical brute force. This limit, in principle, could be overcome by other mixers if they are capable of exploring the underlying problem structure of the COPs. That would explain the aforementioned numerical experiments since, while in small solutions spaces, the Grover mixer provides good results by the ability to mix quickly, in larger solutions spaces, the quadratic progress would it is not enough to get satisfactory results with a small number of layers, drastically compromising algorithm performance.

That idea is strongly endorsed by the numerical study of Golden et al. [11] for the constrained problems Max k-Vertex Cover, k-Densest Subgraph, and Max Bisection. In particular, for simulations up to 18 qubits, the clique mixer performs exponentially better than the GM-Th-QAOA concerning the number of layers to achieve a fixed approximation. Furthermore, direct comparisons with Grover's algorithms were realized, with GM-Th-QAOA performing in the same asymptotic scale and, consequently, with clique mixer performing exponentially better.

Other works corroborate the argument that the Grover mixer is limited to the bound of the unstructured search problem. McClean et al. [121] show that the expectation value of QAOA of a single round with Grover mixer evolves at most as Grover's algorithm. Bennett and Wang [23] on the context of QWOA on the complete graph—recall from Sec. 3.10—provides numerical evidence that a phase separation codifying a threshold function with angles choices that emulate Grover's

[16] In Chapter 4, we prove that the expectation value of the variants of QAOA with Grover mixer is invariant over any permutation of states, which means that instances with the same solution space must have the same performance, independently of the structure of the problem.

algorithm provides the maximum amplification of the probability of measuring a given state. In contrast, Benchasattabuse et al. [122] argue that the worst of Grover mixer on the numerical experiments of Golden et al. [17] and Golden et al. [11] could be due to the fact the classical optimizer does not find the optimal angles. In the present work, the main obtained result is the formal proof that the performance of QAOA with the Grover mixer is indeed limited by a quadratic Grover-style speed-up.

Another result that must be commented on is the lower bounds on the performance of GM-QAOA obtained recently by Benchasattabuse et al. [122]. By introducing an original approach based on the use of bounds at the time of the adiabatic evolution associated with the QAOA, Benchasattabuse et al. [122] obtain the minimum number of layers required to achieve a fixed approximation ratio. Adapting to minimization problems, we assume unconstrained optimization problems with non-positive integers costs[17] and define the average (mean), the standard deviation, and minimum solution of the instance as

$$c_{avg} = \frac{1}{M}\sum_{k\in S} c(k),\ c_{sd} = \sqrt{\frac{\sum_{k\in S}(c(k)-c_{avg})^2}{M}},\ c_{min} = \min\{c(k) : k \in S\}, \tag{3.33}$$

respectively[18]. That way, fixing an approximation ratio λ, the Theorem 3 of Benchasattabuse et al. [122] paper gives

$$r \geq \frac{1-|\langle s|\psi^{(r)}\rangle|^2 + c_{avg} - \lambda c_{min}}{4\pi c_{sd}} \geq \frac{c_{avg}-\lambda c_{min}}{4\pi c_{sd}}, \tag{3.34}$$

where $|s\rangle = |+\rangle^{\otimes n}$ and the last inequality follows from $|\langle s|\psi^{(r)}\rangle| \leq 1$.

Benchasattabuse et al. [122] apply the lower bound of Eq. (3.34) for the class of bipartite graphs on the Max-Cut problem. The mean and standard deviation of Max-Cut were analytically computed, resulting in $|\mathcal{E}|/2$ and $\sqrt{|\mathcal{E}|}/2$, respectively. The maximum cut has $|\mathcal{E}|$ edges for bipartite graphs. Since we deal with minimization context, c_{avg} and c_{min} are $-|\mathcal{E}|/2$ and $-|\mathcal{E}|$, respectively. Thus,

$$r \geq \frac{2\lambda-1}{4\pi}\sqrt{|\mathcal{E}|} = \Omega(\sqrt{|\mathcal{E}|}). \tag{3.35}$$

As NISQ devices require a quantum circuit of low depth, Eq. (3.35) indicates a severe limitation of GM-QAOA once to keep the performance asymptotically on the size of the instances, the number of layers cannot be constant, scaling with the square root of the number of edges.

[17] Therefore, unconstrained maximization problems with non-negative integers costs are applicable.

[18] In Chapter 4, we define these statistical quantities using random variables.

Chapter 4

The Statistical Approach

In this chapter, we consider an analysis method for the variants of QAOA with the Grover mixer—GM-QAOA and GM-Th-QAOA—that looks directly into the spectrum of the problem Hamiltonian (i.e., the solution space of the optimization problem). Specifically, we perform the calculations over the probability distribution associated with the Hamiltonian spectrum. Such a statistical approach takes advantage of the property that QAOA with Grover mixer, provided that the initial state is a uniform superposition over the feasible states, is invariant under any permutation of states, allowing abstracting the combinatorial structure of the problem.

More precisely, the Grover mixer operator being invariant under any permutation means that the action operator keeps the same with the relabelling of any pair of states. With the interpretation of QWOA on the complete graph, the intuition about that property is clear since all one-to-one correspondence of vertices on the complete graph is an automorphism [123]. However, we need to prove that the quantity of our interest, the expectation value, is also invariant under any permutation, which is done in Theorem 7.

Theorem 7 *The expectation value $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$ of any QAOA variant with initial state $|s\rangle$ that uses the Grover mixer as the Hamiltonian mixer is invariant over any permutation of states.*

Proof: Defining the permutation operator $U_{j\leftrightarrow k}$ as

$$U_{j\leftrightarrow k} = \mathbb{I} - |j\rangle\langle j| - |k\rangle\langle k| + |k\rangle\langle j| + |j\rangle\langle k|, \tag{4.1}$$

direct calculations using

$$\langle x|s\rangle = \frac{1}{\sqrt{M}}\sum_{y=0}^{M-1}\langle x|y\rangle = \frac{1}{\sqrt{M}} \tag{4.2}$$

and $\langle s|x\rangle = 1/\sqrt{M}$ lead to $U_{j\leftrightarrow k}H_M U_{j\leftrightarrow k} = H_M$ and therefore H_M is invariant under any relabelling of a pair of states. The initial state $|s\rangle$ have the same property since $U_{j\leftrightarrow k}|s\rangle = |s\rangle$. On the other hand, the phase separation does not have that property in general. Despite this, it is transmitted to the expectation value of the QAOA. To show it, we denote by U_p an arbitrary sequence of permutation operators. Let $E_r^{(p)}$ be the expectation value of the operator $U_p^\dagger H_C U_p$. Then, as U_p is unitary and Hermitian, using the property $e^{U^\dagger H U} = U^\dagger e^H U$, which follows from

$$e^{U^\dagger H U} = \sum_{k=0}^{\infty} \frac{(U^\dagger H U)^k}{k!} = \sum_{k=0}^{\infty} \frac{U^\dagger H^k U}{k!} = U^\dagger e^H U, \tag{4.3}$$

[1] we get

$$\begin{aligned}
E_r^{(p)} &= \langle s| \left(\prod_{j=r}^{1} e^{i\gamma_j U_p^\dagger H_Q U_p} e^{i\beta_j H_M} \right) U_p^\dagger H_C U_p \left(\prod_{j=1}^{r} e^{-i\beta_j H_M} e^{-i\gamma_j U_p^\dagger H_Q U_p} \right) |s\rangle \\
&= \langle s| \left(\prod_{j=r}^{1} U_p^\dagger e^{i\gamma_j H_Q} U_p e^{i\beta_j H_M} \right) U_p^\dagger H_C U_p \left(\prod_{j=1}^{r} e^{-i\beta_j H_M} U_p^\dagger e^{-i\gamma_j H_Q} U_p \right) |s\rangle \\
&= \langle s| U_p^\dagger \left(\prod_{j=r}^{1} e^{i\gamma_j H_Q} U_p e^{i\beta_j H_M} U_p^\dagger \right) H_C \left(\prod_{j=1}^{r} U_p e^{-i\beta_j H_M} U_p^\dagger e^{-i\gamma_j H_Q} \right) U_p |s\rangle \\
&= \langle s| \left(\prod_{j=r}^{1} e^{i\gamma_j H_Q} e^{i\beta_j U_p H_M U_p^\dagger} \right) H_C \left(\prod_{j=1}^{r} e^{-i\beta_j U_p H_M U_p^\dagger} e^{-i\gamma_j H_Q} \right) |s\rangle \\
&= \langle s| \left(\prod_{j=r}^{1} e^{i\gamma_j H_C} e^{i\beta_j H_M} \right) H_Q \left(\prod_{j=1}^{r} e^{-i\beta_j H_M} e^{-i\gamma_j H_Q} \right) |s\rangle,
\end{aligned} \tag{4.4}$$

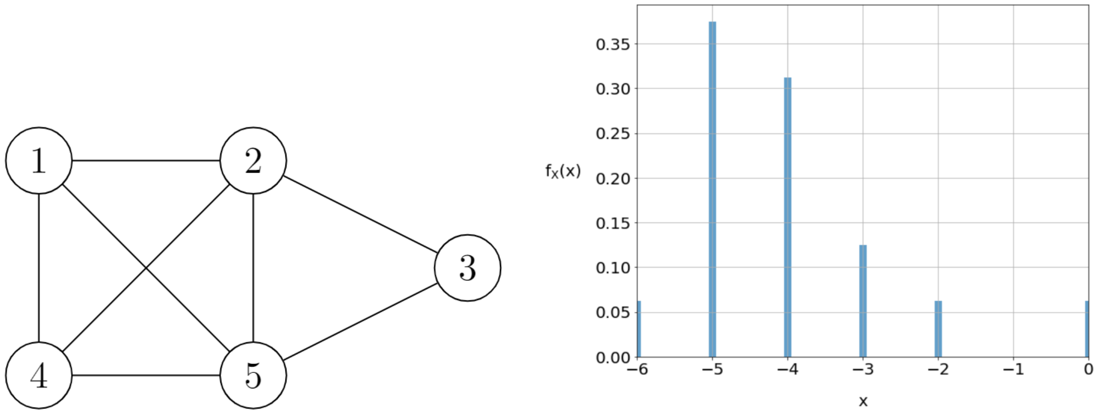
which is the original expectation value, as claimed. □

Note that the theorem is applicable to both GM-QAOA and GM-Th-QAOA since their initial state are $|s\rangle$. The main consequence of Theorem 7 is that two problem Hamiltonians that share the same spectrums must share the same expectation value. That implies that QAOA with Grover mixer on the condition of Theorem 7 is independent of the problem structure, and then, all information required for analysis is provided by the probability distribution. Furthermore, if a given type of instance of a problem converges asymptotically toward a fixed distribution, the result is independent of the size of the instance, avoiding the problem of the barren plateaus.

The mathematical modeling of our analysis is done by using random variables. For a given instance on GM-QAOA or GM-Th-QAOA, let X be the random variable of uniformly sampling an element on the set S and calculating the cost function. The function $f_X(x) = |\{k \in S : c(k) = x\}|/M$ is the probability mass function of X, and the support R_X of X is a countable subset of real numbers. We denote the mean and standard deviation of X by $\mu = \mathrm{E}[X]$ and $\sigma = \sqrt{\mathrm{E}[X-\mu]^2}$, respectively, and

[1] For non-commutative objects such as matrices, we use the notation convention $\sum_{j=a}^{b} x_j = x_b x_{b-1} \ldots x_{a+1} x_a$.

assume $0 < \sigma < \infty$—ignoring the degenerate distribution and taking distributions with finite expectation and standard deviation. Provided that $R_X^{min} \neq 0$ and $|R_X^{min}| < \infty$, the approximation ratio, from Eq. (3.1), can be written in terms of X as $\lambda = E_r(\boldsymbol{\beta}, \boldsymbol{\gamma})/R_X^{min}$ for GM-QAOA and $\lambda = E_r(t)/R_X^{min}$ for GM-Th-QAOA. We also define the random variables Y and Z such that $X = Y + \mu$ and Z is the standard random variable associated with X. Fig. 4.1 shows an example of the probability distribution for an instance of the Max-Cut problem.



(a)

(b)

Figure 4.1: (a) Graph of an instance of Max-Cut problem with 5 vertices and $M = 32$. (b) Probability mass function $f_X(x)$ associated with the instance of (a). We multiply the solutions by -1 to convert it into a minimization problem. The optimal solution has cost $R_X^{min} = -6$ and is given by the partition of the vertices into $\{1, 3, 4\}$ and $\{2, 5\}$. The maximum solution is given by $R_X^{max} = 0$, with trivial partition into $\{1, 2, 3, 4, 5\}$ and $\varnothing$. Note that all partitions are duplicates on the binary codification of the problem.

Before proceeding to our results, note that if $M < \infty$, we cannot consider any probability mass function as $f_X(x)$. This happens because $f_X(x)$ with finite M have rational codomain since the value $f_X(x)$ of an arbitrary $x \in R_X$ is restricted to assume k/M, where k is an integer between 1 and $M - 1$. Then, functions with irrational ranges, as Bernoulli with general p, cannot be analyzed. To bypass this issue, we must assume that $M \to \infty$ and conclude that any real-valued pmf $f_X(x)$ can be obtained in that limit. The proof of that claim follows immediately by a well-known mathematical result that for any $a \in \mathbb{R}$, there is some rational sequence that converges to a on the large limit.

Although $M \to \infty$ is not a real scenario, it is useful for obtaining theoretical results. Furthermore, in some situations, it is convenient for X to be continuous as an asymptotic approximation, either for the theoretical reason of making the cdf continuous or for the practical reason of studying particular continuous distributions, such as the normal distribution. In that approach, all the summations presented from now on are replaced by the respective integrals, with the probability mass function giving way to the probability density function. Both assumptions are quite reasonable in the QAOA context since its main target is NP-Hard optimization problems in which the number of solutions grows exponentially with the size of the entry in such a way that the infinite size limit provides good asymptotic approximations.

Furthermore, although QAOA primarily must be proper to NISQ devices in such a way that the number of layers must be low, in that work, we consider the asymptotic limit of r to several analytical results of this chapter and on Chapter 6, as well as we simulate large values of r for the numerical experiments of Chapter 5. The main reason for this study is to consider the question discussed in Sub. 3.11 of deciding if QAOA with the Grover mixer is limited to a quadratic Grover-like speed-up over classical brute force, addressed in Sec. 4.2 for GM-Th-QAOA and in Chapter 6 for a more general context of QAOA with Grover mixer.

4.1 GM-QAOA analysis

Over this section, we provide explicit expressions for the expectation value of GM-QAOA with statistical quantities of the random variables X, Y, and Z, beginning with $r = 1$ and generalizing for an arbitrary number of layers. Applied to a COP in which the probability distribution is known or even can be approximated, these expressions allow us to obtain analytically the variational parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and compute the expectation value directly from sampling the quantum circuit, avoiding the costly outer loop optimization procedure. Additionally, in cases where the distribution is just partially known, the analytical expression could be used to propose a heuristic method of parameter initialization that at least would be better than the random initialization.

Before enunciating and proving our results, we must prove Lemma 1 and establish the impact on the expectation value of changing the random variable from X to Y.

Lemma 1 *Let H_{C_Y} be the problem Hamiltonian associated with the random variable Y. Then, the expectation value of GM-QAOA can be written as $E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu + \langle\psi^{(r)}|H_{C_Y}|\psi^{(r)}\rangle$.*

Proof: The problem Hamiltonian associated with Y is given by $H_{C_Y} = H_C - \mu\mathbb{I}$. The application of it phase separation, denoted $U_{P_Y}(\gamma)$, is equal to the $U_P(\gamma)$ up to a global phase since

$$U_{P_Y}(\gamma) = e^{-i\gamma H_{C_Y}} = e^{-i\gamma(H_C-\mu\mathbb{I})} = e^{i\gamma\mu\mathbb{I}}e^{-i\gamma H_C} = e^{i\gamma\mu}e^{-i\gamma H_C}. \tag{4.5}$$

Therefore, we can rewrite the expectation value $E_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ as

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \langle\psi^{(r)}|H_C|\psi^{(r)}\rangle = \langle\psi^{(r)}|(H_{C_Y} + \mu\mathbb{I})|\psi^{(r)}\rangle \\ &= \langle\psi^{(r)}|H_{C_Y}|\psi^{(r)}\rangle + \langle\psi^{(r)}|\mu\mathbb{I}|\psi^{(r)}\rangle = \mu + \langle\psi^{(r)}|H_{C_Y}|\psi^{(r)}\rangle. \end{aligned} \tag{4.6}$$

□

4.1.1 Depth 1

For one layer, Theorems 8 and 9 provide explicit expressions in terms of the mean of X, as well as the characteristic function of Y and its derivative.

Theorem 8 *The expectation value of GM-QAOA for a single round is given by*

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu + 2\,\mathrm{Im}\{B(\beta)\varphi_Y^*(\gamma)\varphi_Y'(\gamma)\}. \tag{4.7}$$

Proof: For $r = 1$, the final state of GM-QAOA is given by

$$\begin{aligned} |\psi^{(1)}\rangle &= U_M(\beta)U_P(\gamma)|s\rangle = (\mathbb{I} + B(\beta)|s\rangle\langle s|)U_P(\gamma)|s\rangle \\ &= U_P(\gamma)|s\rangle + B(\beta)\langle s|U_P(\gamma)|s\rangle|s\rangle. \end{aligned} \tag{4.8}$$

The quantity $\langle s|U_P(\gamma)|s\rangle$ is explicitly given by

$$\begin{aligned} \langle s|U_P(\gamma)|s\rangle &= \langle s|\left(\frac{1}{\sqrt{M}}\sum_{k\in S} e^{-i\gamma c(k)}|k\rangle\right) = \frac{1}{M}\sum_{k\in S}\sum_{j\in S} e^{-i\gamma c(k)}\langle j|k\rangle \\ &= \frac{1}{M}\sum_{k\in S} e^{-i\gamma c(k)}. \end{aligned} \tag{4.9}$$

That summation can be equivalently performed by counting the number of solutions such that $c(k) = x$ for each possible value's cost, $x \in R_X$, i.e., $Mf_X(x)$. In that way, we link the characteristic function of X with argument γ by

$$\varphi_X^*(\gamma) = \langle s|U_P(\gamma)|s\rangle = \sum_{x\in R_X} f_X(x)e^{-i\gamma x}, \tag{4.10}$$

and then

$$|\psi^{(1)}\rangle = U_P(\gamma)|s\rangle + B(\beta)\varphi_X^*(\gamma)|s\rangle = (U_P(\gamma) + B(\beta)\varphi_X^*(\gamma)\mathbb{I})|s\rangle. \tag{4.11}$$

The expectation value $E_1(\boldsymbol{\beta},\boldsymbol{\gamma})$ is given by

$$\begin{aligned} E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \langle s|(U_P^\dagger(\gamma) + B^*(\beta)\varphi_X(\gamma)\mathbb{I})H_C(U_P(\gamma) + B(\beta)\varphi_X^*(\gamma)\mathbb{I})|s\rangle \\ &= \langle s|U_P^\dagger(\gamma)H_CU_P(\gamma)|s\rangle + \langle s|U_P^\dagger(\gamma)H_CB(\beta)\varphi_X^*(\gamma)|s\rangle \\ &+ \langle s|B^*(\beta)\varphi_X(\gamma)H_CU_P(\gamma)|s\rangle + \langle s|B^*(\beta)\varphi_X(\gamma)H_CB(\beta)\varphi_X^*(\gamma)|s\rangle. \end{aligned} \tag{4.12}$$

As H_C and $U_P(\gamma)$ are diagonal operators and $|s\rangle$ is an uniform superposition, we have $\langle s|U_P^\dagger(\gamma)H_C U_P(\gamma)|s\rangle = \langle s|H_C|s\rangle = \mu$. Furthermore, similarly to $\langle s|U_P(\gamma)|s\rangle$, $\langle s|H_C U_P(\gamma)|s\rangle$ can be expressed as

$$\langle s|H_C U_P(\gamma)|s\rangle = \frac{1}{M}\sum_{k\in S} c(k)e^{-i\gamma c(k)} = \sum_{x\in R_X} x f_X(x)e^{-i\gamma x}. \tag{4.13}$$

To write Eq. (4.13) in terms of a statistical quantity, note that by Eq. (2.88), $\langle s|H_C U_P(\gamma)|s\rangle^* = \langle s|H_C U_P^\dagger(\gamma)|s\rangle = -i\varphi_Y'(\gamma)$. With these considerations, and using the properties of complex numbers, $|z|^2 = zz^*$, $z+z^* = 2\,\mathrm{Re}\{z\}$, and $\mathrm{Im}\{z\} = \mathrm{Re}\{-iz\}$, we have

$$\begin{aligned} E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mu - iB(\beta)\varphi_X^*(\gamma)\varphi_X'(\gamma) + iB^*(\beta)\varphi_X(\gamma)\varphi_X'^*(\gamma) \\ &\quad + B(\beta)\varphi_X^*(\gamma)B^*(\beta)\varphi_X(\gamma)\mu \\ &= \mu + |B(\beta)|^2|\varphi_X(\gamma)|^2\mu + 2\,\mathrm{Im}\{B(\beta)\varphi_X^*(\gamma)\varphi_X'(\gamma)\}. \end{aligned} \tag{4.14}$$

To finish, with Lemma 1, since the mean of Y is 0, we can eliminate one term of the expression and get Eq. (4.7).

□

To simplify the expression of Theorem 8 and facilitate the analytical optimization of particular distributions, in Theorem 9 we reduce the number of optimization parameters to 1 by using calculus arguments to explicitly give the optimal β value as a function of γ. Before proceeding, we introduce the notation $\mathrm{Arg}(z)$ for a complex number z, which represents the phase of z on the interval $(-\pi,\pi]$. Explicitly, can get $\mathrm{Arg}(z)$ for $z = a + ib$ where $a, b \in \mathbb{R}$ by $\mathrm{Arg}(z) = \arctan(b,a)$, since $\arctan(a,b)$ consider the quadrant of the complex plane.

Theorem 9 *The expectation value of GM-QAOA for a single round is given by*

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu - 2|\phi_Y(\gamma)|(1 + \sin(\mathrm{Arg}(\phi_Y(\gamma)))), \tag{4.15}$$

with an associated parameter β optimal for fixed γ given by

$$\beta = \begin{cases} \frac{\pi}{2} + \mathrm{Arg}(\phi_Y(\gamma)), & \mathrm{Arg}(\phi_Y(\gamma)) \le \frac{\pi}{2} \\ -\frac{3\pi}{2} + \mathrm{Arg}(\phi_Y(\gamma)), & \mathrm{Arg}(\phi_Y(\gamma)) > \frac{\pi}{2}, \end{cases} \tag{4.16}$$

where $\phi_Y(\gamma) = \varphi_Y^*(\gamma)\varphi_Y'(\gamma)$.

Proof: By Eq. (4.7),

$$\begin{aligned} E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mu + 2\,\mathrm{Im}\{[(\cos(\beta) - 1) - i\sin(\beta)][\mathrm{Re}\{\phi_Y(\gamma)\} + i\,\mathrm{Im}\{\phi_Y(\gamma)\}]\} \\ &= \mu + 2(\cos(\beta) - 1)\,\mathrm{Im}\{\phi_Y(\gamma)\} - 2\sin(\beta)\,\mathrm{Re}\{\phi_Y(\gamma)\}. \end{aligned} \tag{4.17}$$

As $\phi_Y(\gamma)$ for given γ is a complex number, we can write it by $|\phi_Y(\gamma)|e^{i\theta}$, where $\theta = \text{Arg}(\phi_Y(\gamma))$. Replacing it in Eq. (4.17), we have

$$\begin{aligned} E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \mu + 2(\cos(\beta) - 1)|\phi_Y(\gamma)|\sin(\theta) - 2\sin(\beta)|\phi_Y(\gamma)|\cos(\theta) \\ &= \mu - 2|\phi_Y(\gamma)|(\sin(\beta)\cos(\theta) - \cos(\beta)\sin(\theta) + \sin(\theta)) \\ &= \mu - 2|\phi_Y(\gamma)|(\sin(\beta - \theta) + \sin(\theta)), \end{aligned} \tag{4.18}$$

where in the last equality we use the trigonometric identity $\sin(x - y) = \sin x \cos y - \cos x \sin y$. As the factor $2|\phi_Y(\gamma)|$ is non-negative, the optimal parameter for β for a fixed γ, is totally determined by θ, which is, in turn, dependent on γ. Finding the minimum of $E_1(\boldsymbol{\beta},\boldsymbol{\gamma})$ in that condition is equivalent to finding the maximum of $\sin(\beta - \theta) + \sin(\theta)$. Taking the first and the second partial derivatives with respect to β,

$$\begin{aligned} &\frac{\partial}{\partial\beta}[\sin(\beta-\theta) + \sin(\theta)], = \cos(\beta - \theta) \\ &\frac{\partial^2}{\partial\beta^2}[\sin(\beta-\theta) + \sin(\theta)] = -\sin(\beta - \theta). \end{aligned} \tag{4.19}$$

With the first derivative, the extreme points are the solutions of the equation $\cos(\beta - \theta) = 0$, that is, $\beta = \pi(n - 1/2) + \theta,\ \forall n \in \mathbb{Z}$. Conversely, by the second derivative, $n = 1$ gives a maximum point because $-\sin(\pi/2 + \theta - \theta) = -1$ is negative. Therefore, $\beta = \pi/2 + \theta$ is optimal for the given γ. Replacing it in Eq. (4.18) we get Eq. (4.15). For the optimal value of β, note that $\beta \in (-\pi, \pi]$ and $\text{Arg}(\phi_Y(\gamma)) \in (-\pi, \pi]$ and therefore we have to adjust $\beta = \pi/2 + \theta$ to Eq. (4.16).

□

A particular case, given by Corollary 1, is when the distribution $f_Y(x)$ is symmetric, or equivalently, $f_X(x)$ is symmetric around the mean.

Corollary 1 *For an instance of GM-QAOA with a random variable X such that $f_Y(x)$ is a symmetric distribution, the expectation value for a single round is given by*

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu - 2|\phi_Y(\gamma)|, \tag{4.20}$$

with an associated parameter β optimal for fixed γ given by

$$\beta = \begin{cases} \frac{\pi}{2}, & Arg\,(\phi_Y(\gamma)) = 0 \\ -\frac{\pi}{2}, & Arg\,(\phi_Y(\gamma)) = \pi. \end{cases} \tag{4.21}$$

Proof: With the well-known property that the Fourier transform of an even and an odd function are a real-valued function and a purely imaginary function, respectively, as $f_Y(x)$ is even and $xf_Y(x)$ odd,

$$\begin{aligned}\phi_Y(\gamma) &= \left(\sum_{x\in R_Y} f_Y(x)e^{-i\gamma x}\right)\left(i\sum_{x\in R_Y} xf_Y(x)e^{i\gamma x}\right)\\ &= -\left(\sum_{x\in R_Y} f_Y(x)\cos(\gamma x)\right)\left(\sum_{x\in R_Y} xf_Y(x)\sin(\gamma x)\right),\end{aligned} \tag{4.22}$$

which is a real-valued function. The phase of $\phi_Y(\gamma)$ is 0 or π. Both gives from Eq. (4.15), Eq. (4.20). Replacing each one in Eq. (4.16) we get Eq. (4.21).

□

Another consequence in symmetric around the mean distributions is that since $\varphi_Y^*(\gamma)$ and $\varphi'_Y(\gamma)$ are real-valued functions in that case, we can rewrite Eq. (4.7) as

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu + 2\varphi_Y(\gamma)\varphi'_Y(\gamma)\,\mathrm{Im}\{B(\beta)\}. \tag{4.23}$$

Now, note that the product $\varphi_Y(\gamma)\varphi'_Y(\gamma)$ is odd function on γ—since $\varphi_Y(\gamma)$ and $\varphi'_Y(\gamma)$ are even and odd functions on γ, respectively—and $\mathrm{Im}\{B(\beta)\}$ is a odd function on β. Thus, the landscape of GM-QAOA is symmetric for diagonally opposite quadrants on the range $\beta \in (-\pi,\pi]$ and $\gamma \in (-x,x]$ for some $0 < x < \infty$. Consequently, we must have two symmetric global minima in diagonally opposite quadrants and two symmetric global maxima in the other diagonally opposite quadrants. An example can be found in Fig. 5.2, on Chapter 5.

4.1.2 Arbitrary depth

Theorem 10 generalizes the result of Theorem 8 for arbitrary r, providing an expectation value expression with the same statistical quantities. For reading fluidity purposes, the proof of the theorem is shown before its statement so that the notation is introduced naturally.

For arbitrary r, the final state of GM-QAOA can be written as

$$|\psi^{(r)}\rangle = \prod_{j=1}^{r} U_M(\beta_j)U_P(\gamma_j)|s\rangle = \prod_{j=1}^{r}(\mathbb{I} + B(\beta_j)|s\rangle\langle s|)U_P(\gamma_j)|s\rangle. \tag{4.24}$$

Representing $U_P(\gamma_j)$ as the sum of projections and introducing the r-dimensional vector $\boldsymbol{z} = (z_1,\ldots,z_r)$, where each element runs over the set of M feasible solutions of S, we can rewrite Eq. (4.24) as

$$|\psi^{(r)}\rangle = \left(\sum_{\boldsymbol{z}}\prod_{j=1}^{r}(\mathbb{I} + B(\beta_j)|s\rangle\langle s|)e^{-i\gamma_j c(z_j)}|z_j\rangle\langle z_j|\right)|s\rangle. \tag{4.25}$$

The expression can be written in an equivalent manner using the r-bit string $\boldsymbol{x} =$

$(x_1, \ldots, x_r)$ in such a way that

$$|\psi^{(r)}\rangle = \left(\sum_{\boldsymbol{x}} \sum_{\boldsymbol{z}} \prod_{j=1}^{r} (B(\beta_j)|s\rangle\langle s|)^{x_j} e^{-i\gamma_j c(z_j)} |z_j\rangle\langle z_j| \right) |s\rangle. \tag{4.26}$$

The string $\boldsymbol{x}$ acts like an incident vector to the presence of the factor $B(\beta_j)|s\rangle\langle s|$ for each index j. If $x_j = 0$, $(B(\beta_j)|s\rangle\langle s|)^{x_j} = \mathbb{I}$ and if $x_j = 1$, $(B(\beta_j)|s\rangle\langle s|)^{x_j} = B(\beta_j)|s\rangle\langle s|$. Using the projection representation of problem Hamiltonian and superscripts L and R replacing the original $\boldsymbol{x}$ and $\boldsymbol{z}$ to differ the left (bra) and right (ket) part of the expression, respectively, we express the expectation value as

$$\begin{gathered} E_r(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle s| \left(\sum_{\boldsymbol{x}^{(L)}} \sum_{\boldsymbol{z}^{(L)}} \prod_{j=r}^{1} |z_j^{(L)}\rangle\langle z_j^{(L)}| e^{i\gamma_j c(z_j^{(L)})} (B^*(\beta_j)|s\rangle\langle s|)^{x_j^{(L)}} \right) \\ \left(\sum_z c(z)|z\rangle\langle z| \right) \left(\sum_{\boldsymbol{x}^{(R)}} \sum_{\boldsymbol{z}^{(R)}} \prod_{j=1}^{r} (B(\beta_j)|s\rangle\langle s|)^{x_j^{(R)}} e^{-i\gamma_j c(z_j^{(R)})} |z_j^{(R)}\rangle\langle z_j^{(R)}| \right) |s\rangle. \end{gathered} \tag{4.27}$$

Rearranging the factors,

$$\begin{gathered} E_r(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \langle s| \left(\sum_{\boldsymbol{x}^{(L)}} \sum_{\boldsymbol{z}^{(L)}} e^{\sum_{j=1}^{r} i\gamma_j c(z_j^{(L)})} \left[\prod_{j: x_j^{(L)}=1} B^*(\beta_j) \right] \right. \\ \left. \left[\prod_{j=r}^{1} |z_j^{(L)}\rangle\langle z_j^{(L)}| (|s\rangle\langle s|)^{x_j^{(L)}} \right] \right) \left(\sum_z c(z)|z\rangle\langle z| \right) \left(\sum_{\boldsymbol{x}^{(R)}} \sum_{\boldsymbol{z}^{(R)}} e^{\sum_{j=1}^{r} -i\gamma_j c(z_j^{(R)})} \right. \\ \left. \left[\prod_{j: x_j^{(R)}=1} B(\beta_j) \right] \left[\prod_{j=1}^{r} (|s\rangle\langle s|)^{x_j^{(R)}} |z_j^{(R)}\rangle\langle z_j^{(R)}| \right] \right) |s\rangle. \end{gathered} \tag{4.28}$$

With $\langle s|z_1^{(L)}\rangle = \langle z_1^{(R)}|s\rangle = 1/\sqrt{M}$, we can write Eq. (4.28) as

$$\begin{gathered} E_r(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{\boldsymbol{x}^{(L)}, \boldsymbol{x}^{(R)}} \sum_{z} \sum_{\boldsymbol{z}^{(L)}, \boldsymbol{z}^{(R)}} e^{\sum_{j=1}^{r} i\gamma_j c(z_j^{(L)})} e^{\sum_{j=1}^{r} -i\gamma_j c(z_j^{(R)})} \left[\prod_{j: x_j^{(L)}=1} B^*(\beta_j) \right] \\ \left[\prod_{j: x_j^{(R)}=1} B(\beta_j) \right] \left[\langle z_1^{(L)}| (|s\rangle\langle s|)^{x_1^{(L)}} \right] \left[\prod_{j=r}^{2} |z_j^{(L)}\rangle\langle z_j^{(L)}| (|s\rangle\langle s|)^{x_j^{(L)}} \right] \\ \left(\frac{1}{M} c(z)|z\rangle\langle z| \right) \left[\prod_{j=2}^{r} (|s\rangle\langle s|)^{x_j^{(R)}} |z_j^{(R)}\rangle\langle z_j^{(R)}| \right] \left[(|s\rangle\langle s|)^{x_1^{(R)}} |z_1^{(R)}\rangle \right]. \end{gathered} \tag{4.29}$$

To proceed, consider two situations on the left side portion of the expression. Firstly, if we have a sequence of adjacent indices in left side of expression such that $x_k^{(L)} = x_{k+1}^{(L)} = \ldots = x_{k+l-1}^{(L)} = x_{k+l}^{(L)} = 0$ for arbitrary $k \geq 2$ and $l \geq k$, then the product involving such indices results

$$\prod_{j=k+l}^{k} |z_j^{(L)}\rangle\langle z_j^{(L)}|(|s\rangle\langle s|)^{x_j^{(L)}} = \left(\prod_{j=k}^{k+l-1} \delta(z_j^{(L)}, z_{j+1}^{(L)})\right) |z_k^{(L)}\rangle\langle z_{k+l}^{(L)}|, \tag{4.30}$$

where $\delta(x,y)$ denotes the Kronecker delta. Next, if $x_k^{(L)} = 1$ for some $k \geq 2$, then for the indices k and $k+1$ we have

$$\prod_{j=k+1}^{k} |z_j^{(L)}\rangle\langle z_j^{(L)}|(|s\rangle\langle s|)^{x_j^{(L)}} = \frac{1}{M}|z_k^{(L)}\rangle\langle z_{k+1}^{(L)}|(|s\rangle\langle s|)^{x_{k+1}^{(L)}}. \tag{4.31}$$

That is, the projector $|s\rangle\langle s|$ breaks the chain of Kronecker delta factors with $\langle z_k^{(L)}|s\rangle = \langle s|z_{k+1}^{(L)}\rangle = 1/\sqrt{M}$.

Combining both situations in a sequence of $x_k^{(L)} = x_{k+1}^{(R)} = \ldots = x_{k+l-2}^{(L)} = x_{k+l-1}^{(L)} = 0$ and $x_{k+l}^{(L)} = 1$ for the indices k up to $k+l+1$, the product is given by

$$\begin{aligned}\prod_{j=k+l+1}^{k} |z_j^{(L)}\rangle\langle z_j^{(L)}|(|s\rangle\langle s|)^{x_j^{(L)}} = &\left(\frac{1}{M}\prod_{j=k}^{k+l-1} \delta(z_j^{(L)}, z_{j+1}^{(L)})\right) \\ &|z_k^{(L)}\rangle\langle z_{k+l+1}^{(L)}|(|s\rangle\langle s|)^{x_{k+l+1}^{(L)}}.\end{aligned} \tag{4.32}$$

Since the ket $|z_1^{(L)}\rangle$ is already cancel in Eq. (4.29), we can apply recursively the above process for all elements of the set

$$\begin{aligned}\mathcal{P}_L = \{\{k, k+1, \ldots, k+l-1, k+j\} :& x_{k-1}^{(L)} = x_{k+j}^{(L)} = 1, \\ & x_j^{(L)} = 0 \ \forall k-1 < j < k+l\},\end{aligned} \tag{4.33}$$

where we conveniently set $x_0^{(L)} = 1$. Let L^{max} be the biggest index j such that $x_j^{(L)} = 1$. Note that the indices j such that $j > L^{max}$ are not considered on the set $\mathcal{P}_L$. For them, we define $\mathcal{P}_{0L} = \{j : j > L^{max}\}$. If $L^{max} = r$, then the ket $|z\rangle$ is break with $\langle s|z\rangle = 1/\sqrt{M}$. Otherwise, we have $\delta(z_r^{(L)}, z)$. Thus, we can extend z as the $(r+1)$th index of the right size with $z_{r+1}^{(L)} = z$.

Following analog arguments we set for the right side of the expression

$$\begin{aligned}\mathcal{P}_R = \{\{k, k+1, \ldots, k+l-1, k+j\} :& x_{k-1}^{(R)} = x_{k+j}^{(R)} = 1, \\ & x_j^{(R)} = 0 \ \forall k-1 < j < k+l\},\end{aligned} \tag{4.34}$$

with $x_0^{(R)} = 1$, $z_{r+1}^{(R)} = z$, $\mathcal{P}_{0R} = \{j : j > R^{max}\}$, and R^{max} being the biggest index j such that $x_j^{(R)} = 1$.

Therefore, we can express Eq. (4.29) as

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = &\sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} \sum_{z} \sum_{\boldsymbol{z}^{(L)},\boldsymbol{z}^{(R)}} e^{\sum_{j=1}^{r} i\gamma_j c(z_j^{(L)})} e^{\sum_{j=1}^{r} -i\gamma_j c(z_j^{(R)})} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right] \\ &\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_L}\left(\frac{1}{M}\prod_{j\in\mathcal{P}\smallsetminus\{k+l\}} \delta(z_j^{(L)},z_{j+1}^{(L)})\right)\right] \\ &\left(\prod_{j\in\mathcal{P}_{0L}} \delta(z_j^{(L)},z_{j+1}^{(L)})\right)\left(\frac{1}{M}c(z)\right)\left(\prod_{j\in\mathcal{P}_{0R}} \delta(z_{j+1}^{(R)},z_j^{(R)})\right) \\ &\left[\prod_{\mathcal{P}\in\mathcal{P}_R}\left(\frac{1}{M}\prod_{j\in\mathcal{P}\smallsetminus\{k+l\}} \delta(z_{j+1}^{(R)},z_j^{(R)})\right)\right]. \end{aligned} \tag{4.35}$$

Note that for a given element of $\mathcal{P}_L$, the product

$$\frac{1}{M}\prod_{j\in\mathcal{P}\smallsetminus\{k+l\}} \delta(z_j^{(L)},z_{j+1}^{(L)}) \tag{4.36}$$

is equal to $1/M$ if $z_k^{(L)} = z_{k+1}^{(L)} = \ldots = z_{k+l-1}^{(L)} = z_{k+l}^{(L)}$, and 0 otherwise. That way, we can condensate variables $z_k^{(L)}, z_{k+1}^{(L)}, \ldots, z_{k+l-1}^{(L)}, z_{k+l}^{(L)}$ in a single variable. The same can be done for the set $\mathcal{P}_L$. By an analogous argument, for the product

$$\left(\prod_{j\in\mathcal{P}_{0L}} \delta(z_j^{(L)},z_{j+1}^{(L)})\right)\left(\frac{1}{M}c(z)\right)\left(\prod_{j\in\mathcal{P}_{0R}} \delta(z_j^{(R)},z_{j+1}^{(R)})\right), \tag{4.37}$$

we can combine the variables $z_{L^{max}+1}^{(L)}, z_{L^{max}+2}^{(L)}, \ldots, z_{r-1}^{(L)}, z_r^{(L)}$, z, $z_{R^{max}+1}^{(R)}, z_{R^{max}+2}^{(R)}, \ldots, z_{r-1}^{(R)}, z_r^{(R)}$. Then, follows

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = &\sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right]\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right] \\ &\left[\prod_{\mathcal{P}\in\mathcal{P}_L}\frac{1}{M}\sum_z \exp\left(\sum_{j\in\mathcal{P}} i\gamma_j c(z)\right)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_R}\frac{1}{M}\sum_z \exp\left(\sum_{j\in\mathcal{P}} -i\gamma_j c(z)\right)\right] \\ &\left[\frac{1}{M}\sum_z c(z)\exp\left(\left(\sum_{j\in\mathcal{P}_{0L}} i\gamma_j c(z)\right)+\left(\sum_{j\in\mathcal{P}_{0R}} -i\gamma_j c(z)\right)\right)\right]. \end{aligned} \tag{4.38}$$

The factor involving the sets $\mathcal{P}_{0L}$ and $\mathcal{P}_{0R}$ can be simplified by noting that for a j such that $j \in Q_{0L}$ and $j \in Q_{0R}$, $\gamma_j - \gamma_j = 0$. To consider all simplification, we define the set $\mathcal{P}_0$ as $\{-j : L^{max} \geq j > R^{max}\}$ if $L^{max} > R^{max}$, $\{j : R^{max} \geq j > L^{max}\}$ if $L^{max} < R^{max}$, and null if $L^{max} = R^{max}$. Thus,

$$\left(\sum_{j\in\mathcal{P}_{0L}} i\gamma_j c(z)\right)+\left(\sum_{j\in\mathcal{P}_{0R}} -i\gamma_j c(z)\right) = \sum_{j\in\mathcal{P}_0} i\gamma_j c(z). \tag{4.39}$$

Moreover, we can replace the characteristic functions and their derivatives in the expression in an analogous way to depth 1 analysis, resulting in

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = -i \sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j) \right] \left[\prod_{j:x_j^{(R)}=1} B(\beta_j) \right] \\ \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_X \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi_X^* \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\varphi_X' \left(\sum_{j\in\mathcal{P}_0} \gamma_j \right) \right]. \end{aligned} \tag{4.40}$$

For any order pair of $\boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(R)}$ in which $\boldsymbol{x}^{(L)} \neq \boldsymbol{x}^{(R)}$, if $\boldsymbol{x}^{(L)} = \boldsymbol{x}_1$ and $\boldsymbol{x}^{(R)} = \boldsymbol{x}_2$, than the term with $(\boldsymbol{x}_2, \boldsymbol{x}_1)$ is the complex conjugate of the term with $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Therefore, with the property $z + z^* = 2\,\mathrm{Re}\{z\}$ and noticing that if $\boldsymbol{x}^{(L)} = \boldsymbol{x}^{(R)}$, $\mathcal{P}_0 = \varnothing$ and the derivative of characteristic function is reduced to $\varphi_X'(0)$, we get

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = &\left(\sum_{\boldsymbol{x}^{(L)}=\boldsymbol{x}^{(R)}} \varphi_X'(0) \left[\prod_{j:x_j^{(L)}=1} |B(\beta_j)|^2 \right] \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \left| \varphi_X \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right|^2 \right] \right) \\ &+ \left(2\,\mathrm{Im} \left\{ \sum_{\boldsymbol{x}^{(L)}<\boldsymbol{x}^{(R)}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j) \right] \left[\prod_{j:x_j^{(R)}=1} B(\beta_j) \right] \right. \right. \\ &\left. \left. \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_X \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi_X^* \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\varphi_X' \left(\sum_{j\in\mathcal{P}_0} \gamma_j \right) \right] \right\} \right), \end{aligned} \tag{4.41}$$

where $\boldsymbol{x}^{(L)} < \boldsymbol{x}^{(R)}$ is an abuse of notation to compare the respective numbers of the binary representation of $\boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(R)}$. As $\varphi_X'(0) = -i\mu$, by Lemma 1, we can cancel the first main summation of Eq. (4.41) and restrict the second summation to $L^{max} < R^{max}$ terms. With this, we prove Theorem 10.

Theorem 10 *The expectation value of GM-QAOA for an arbitrary number of rounds r is given by*

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu + 2\,\mathrm{Im} \left\{ \sum_{\boldsymbol{x}^{(L)}<\boldsymbol{x}^{(R)}:L^{max}<R^{max}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j) \right] \left[\prod_{j:x_j^{(R)}=1} B(\beta_j) \right] \right. \\ \left. \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_Y \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi_Y^* \left(\sum_{j\in\mathcal{P}} \gamma_j \right) \right] \left[\varphi_Y' \left(\sum_{j\in\mathcal{P}_0} \gamma_j \right) \right] \right\}. \end{aligned} \tag{4.42}$$

Corollary 2 gives the complexity of Eq. (4.42). To prove it, we need to count the number of terms of the main summation of the expression, which is done in Lemma 2. The original summation of Eq. (4.40) has 4^r terms, but even canceling terms on Eq. (4.42), the complexity is still of order $\Theta(4^r)$.

Lemma 2 *The number of terms of the main summation of Eq.* (4.42) *is* $\frac{4^r-1}{3}$.

Proof: We need to count the number of order pair $(\boldsymbol{x}^{(L)}, \boldsymbol{x}^{(R)})$ such that $L^{max} = R^{max}$, subtract it from the original 4^r terms and cut off half of them. There is a single combination in which $L^{max} = 0$ and $4^{L^{max}-1}$ if $L^{max} > 0$. The last case can be seen as the number of combinations of the remainder $L^{max} - 1$ most significant bits of both $\boldsymbol{x}^{(L)}$ and $\boldsymbol{x}^{(R)}$. Therefore, by counting all possibilities of L^{max}, the number of terms, using the summation

$$\sum_{j=0}^{n} x^j = \frac{x^{n+1} - 1}{x - 1}, \tag{4.43}$$

which follows denoting it by S_n and taking,

$$xS_n = x + x^2 + \ldots + x^n + x^{n+1} = S_n - 1 + x^{n+1} \;\Rightarrow\; S_n = \frac{x^{n+1} - 1}{x - 1}, \tag{4.44}$$

is computed by

$$\frac{1}{2}\left(4^r - \left(1 + \sum_{j=1}^{r} 4^{j-1}\right)\right) = \frac{4^r - 1}{3}, \tag{4.45}$$

as desired. □

Corollary 2 *Given attributions for the angles $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, the complexity of the expectation value expression of GM-QAOA, given by Eq.* (4.42)*, depends on the numbers of layers r by $\mathcal{O}(4^r)$.*

Proof: To begin, for a given order pair $(\boldsymbol{x}^{(L)}, \boldsymbol{x}^{(R)})$, the sets $\mathcal{P}_L$, $\mathcal{P}_R$, $\mathcal{P}_0$, as well as the quantities L^{max} and R^{max} can be computed in polynomial time on r. Then, by Lemma 2, the number of terms of the main summation of Eq. (4.42) is $\Theta(4^r)$. For each term, the number of any of the products has at most $2r$ factors, and each characteristic function or derivative has at most r parameters γ_j to add. Therefore, the whole expression has the claimed complexity. □

For $r = 1$, Eq. (4.42) is reduced to Eq. (4.7). That is the only depth in which we can optimize GM-QAOA analytically given a particular distribution (especially with Theorem 9). From 2 layers onwards, we optimize numerically using computational tools. For $r = 2$, the main summation has 5 terms and we can explicitly give the expression using Theorem 10, resulting in

$$\begin{aligned} E_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \mu &+ 2\,\mathrm{Im}\{B(\beta_1)\varphi_Y^*(\gamma_1)\varphi_Y'(\gamma_1)\} \\ &+ 2\,\mathrm{Im}\{B(\beta_2)\varphi_Y^*(\gamma_1 + \gamma_2)\varphi_Y'(\gamma_1 + \gamma_2)\} \\ &+ 2|B(\beta_1)|^2|\varphi_Y(\gamma_1)|^2\,\mathrm{Im}\{B(\beta_2)\varphi_Y^*(\gamma_2)\varphi_Y'(\gamma_2)\} \\ &+ 2\,\mathrm{Im}\{B(\beta_1)B(\beta_2)\varphi_Y^*(\gamma_1)\varphi_Y^*(\gamma_2)\varphi_Y'(\gamma_1 + \gamma_2)\} \\ &+ 2\,\mathrm{Im}\{B^*(\beta_1)B(\beta_2)\varphi_Y(\gamma_1)\varphi_Y^*(\gamma_1 + \gamma_2)\varphi_Y'(\gamma_2)\}. \end{aligned} \tag{4.46}$$

Unfortunately, we cannot keep showing explicit expressions for larger values of r due to the exponential number of terms on r. For instance, the sequence of the number of non-trivial terms of the expectation value for the first 8 depth levels is $1, 5, 21, 85, 341, 1365, 5461, 21845$.

Recall that by introducing the auxiliary random variable Y, we neutralize the impact of the mean of X with the trivial term μ in the expression of expectation value. The analog can be done for the standard deviation by introducing the standard random variable Z with the properties $\varphi_Y(\gamma) = \varphi_Z(\sigma\gamma)$ and $\varphi'_Y(\gamma) = \sigma\varphi'_Z(\sigma\gamma)$, resulting immediately in Corollary 3.

Corollary 3 *The expectation value of GM-QAOA for an arbitrary number of rounds r is given by*

$$\begin{aligned} E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu + 2\sigma \operatorname{Im}\Bigg\{ \sum_{\boldsymbol{x}^{(L)}<\boldsymbol{x}^{(R)}:L^{max}<R^{max}} \Bigg[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\Bigg]\Bigg[\prod_{j:x_j^{(R)}=1} B(\beta_j)\Bigg] \\ \Bigg[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_Z\Big(\sigma\sum_{j\in\mathcal{P}}\gamma_j\Big)\Bigg]\Bigg[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi_Z^*\Big(\sigma\sum_{j\in\mathcal{P}}\gamma_j\Big)\Bigg]\Bigg[\varphi'_Z\Big(\sigma\sum_{j\in\mathcal{P}_0}\gamma_j\Big)\Bigg]\Bigg\}. \end{aligned} \tag{4.47}$$

The corollary implies that $E_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ deviates from the mean proportionally to σ. Therefore, for a given X, it is straightforward to consider the negative of the standard score, i.e., a $C_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ such that $E_r(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mu - C_r(\boldsymbol{\beta},\boldsymbol{\gamma})\sigma$, as a performance metric. To to simplify writing, we call $C_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ by standard score. Furthermore, the angles γ_j are inversely proportional to σ—i.e., it modifies inversely proportional with changes on the scale of the distribution.

The standard score has an advantage over the expectation value and the approximation ratio since it can be used as a comparison between distinct distributions and combinatorial optimization problems. The expectation value can vary greatly between different instances by shifting location and changing the scale, and even the approximation ratio, for instance, is not applicable for distribution with $R_X^{min} \to -\infty$ or $R_X^{min} = 0$. Furthermore, note that GM-QAOA performance is not affected by the two first statistical moments, since shifting location and changing the scale on probability distribution results in the same transformations of the outcome of GM-QAOA. Therefore, the performance depends only on moments of higher order, such as skewness and kurtosis. In particular, if X has all finite moments, from Eq. (2.84) we can expand the characteristic function of Z by

$$\varphi_Z(\omega) = 1 - \frac{\omega^2}{2} - \frac{i\,\mathrm{Skew}[X]\omega^3}{6} + \frac{\mathrm{Kurt}[X]\omega^4}{24} + \sum_{n=5}^{\infty} \frac{i^n\,\mathrm{E}[Z^n]\omega^n}{n!}. \tag{4.48}$$

4.1.3 Upper bounds on the standard score for GM-QAOA

A natural question that arises on Corollary 3 concerns the general upper bounds for $C_r(\boldsymbol{\beta}, \boldsymbol{\gamma})$. We denote the maximum possible value of $C_r(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by $C^{GM}(r)$. Applying the individual bounds $\varphi_Y(\gamma) \leq 1$ and $\varphi'_Y(\gamma) \leq \mathrm{E}[|Y|] \leq \sigma$ on either Eq. (4.7) or Eq. (4.15) gives the bound $C^{GM}(1) \leq 4$. The inequality $\mathrm{E}[|Y|] \leq \sigma$ follows by setting the random variable $|Y|$ and the convex function x^2 on Jensen's inequality. Theorem 11 refine the bound for $C^{GM}(1) \leq \frac{8\sqrt{6}}{9} \approx 2.178$ by using calculus arguments with a bound on the second derivative of the characteristic function.

Theorem 11 *For a single round, the maximum standard score $C_1(\boldsymbol{\beta}, \boldsymbol{\gamma})$ achieved by GM-QAOA is bounded by $C^{GM}(1) \leq \frac{8\sqrt{6}}{9}$.*

Proof: The second derivative of the characteristic function is bounded by $\varphi''_Y(\gamma) \leq \mathrm{E}[|Y|^2] = \sigma^2$. Translated the our three individual bounds into the random variable Z, we have $|\varphi_Z(\sigma\gamma)|, |\varphi'_Z(\sigma\gamma)|, |\varphi''_Z(\sigma\gamma)| \leq 1$. We are interested in bound the quantity $|\varphi^*_Z(\sigma\gamma)\varphi'_Z(\sigma\gamma)|$. To get that, we can think of our problem as the kinematics problem of maximizing the product of the distance and velocity given that both and the acceleration are bounded. As $\varphi_Z(\sigma\gamma)$ is a complex number in the general case, the problem has two dimensions. Do we ask what the maximum scalar distance $|\varphi_Z(\sigma\gamma)|$ for a fixed scalar velocity $|\varphi'_Z(\sigma\gamma)|$ in which we do not violate the maximum distance of 1 unit? Of course, we minimize the distance traveled by assuming the maximum deceleration, which in our case is $|\varphi''_Z(\sigma\gamma)| = 1$. Thus, for a scalar velocity $0 \leq x \leq 1$, the scalar distance traveled until deceleration completely is $x^2/2$, and we must be in a distance of at most $1 - x^2/2$. Therefore,

$$|\varphi^*_Z(\sigma\gamma)\varphi'_Z(\sigma\gamma)| \leq \left(1 - \frac{x^2}{2}\right) x. \tag{4.49}$$

The derivative equal to zero gives the maximum on the point $x = \sqrt{\frac{2}{3}}$, which is evaluated on the bound to $\left(\frac{2}{3}\right)^{3/2}$ and by Eq. (4.15) the claimed bound follows. □

For general r, on the other hand, applying inequalities on $|\varphi_Y(\gamma)|$, $|\varphi'_Y(\gamma)|$ and $|B(\beta)|$ is insufficient to obtain a satisfactory bound since it would grow exponentially on r. Directly from Lemma 2, we see the growth would be at least on 4^r order. Indeed, the growth is of an order of 9^r. We get the exact number in the following way.

Firstly, we count the individual bound over all 4^r terms of Eq. (4.40) and in sequence, we subtract the number of terms canceled on Eq. (4.42), that is, the terms in which $L^{max} = R^{max}$. For each term, the bound is determined exclusively by the number of $B(\beta)$ terms. The bound over Eq. (4.40) gives

$$\sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} \left[\prod_{j:\ x_j^{(L)}=1} 2 \right] \left[\prod_{j:\ x_j^{(R)}=1} 2 \right] = \left(\sum_{j=0}^{r} \binom{r}{j} 2^j \right)^2 = 9^r, \tag{4.50}$$

where the first equality follows from noting that the number of factors of each product on a given term is the Hamming weight of the binary representation of $\boldsymbol{x}^{(L)}$ or $\boldsymbol{x}^{(R)}$, and the last from the expansion of $(1+2)^r$ with the binomial theorem.

To the terms in which $L^{max} = R^{max}$, if $L^{max} = 0$, there is a unique combination that sums 1 unit to the bound. For $L^{max} > 0$, in both $\boldsymbol{x}^{(L)}$ or $\boldsymbol{x}^{(R)}$, the L^{max}th bit is 1. We must count the Hamming weight of the $L^{max} - 1$ remainder most significant bits. Thus, for a given L^{max}, we add

$$\begin{aligned} &\sum_{x=0}^{L^{max}-1} \sum_{y=0}^{L^{max}-1} \binom{L^{max}-1}{x} \binom{L^{max}-1}{y} 2^{2+x+y} \\ &= 4 \left(\sum_{j=0}^{L^{max}-1} \binom{L^{max}-1}{j} 2^j \right)^2 = 4\ 9^{L^{max}-1} \end{aligned} \tag{4.51}$$

units. Considering the subtraction over the initial 9^r and summing over all possible L^{max}, we conclude using the summation of Eq. (4.43) that

$$C^{GM}(r) \leq 9^r - \left(1 + 4 \sum_{j=1}^{r} 9^{j-1} \right) = \frac{9^r - 1}{2}. \tag{4.52}$$

Unfortunately, due to the complexity of the expression of Theorem 10, direct analytical treatment to improve the bound of Eq. (4.52) is unfeasible. This topic is returned on Chapter 6, in which we use an indirect method to bound the expectation value of Grover-based QAOA, the more general version of GM-QAOA.

4.1.4 The binary function

Before proceeding to the analysis of GM-Th-QAOA, we present the binary function, an application of GM-QAOA, which is the core of that analysis. In that work, we define the binary function as a function that assigns value −1 for elements belonging to a subset of marked elements and 0 otherwise. We denote by ρ the ratio of marked elements to the entire domain of the function. Note that the binary function is similar to the function of Eq. (3.2) used on the unstructured search problem but adapted to the context of minimization problems. By the statistical interpretation of optimization problems, the distribution of the binary function follows RBernoulli(ρ).

Note that minimizing GM-QAOA with the binary function of ratio ρ as input is equivalent to minimizing GM-Th-QAOA with a fixed threshold value in which the ratio of states equal or below the threshold is ρ. The operators are identical, and the observable, although different, have the equivalent goal of maximizing the probability of measuring a set of states with ratio ρ. Therefore, we can use Golden et al. [26] result on $r = 1$ of GM-Th-QAOA, discussed on Subsec. 3.8.2. That way, if $\rho > 0.25$ we have $E_1(\boldsymbol{\beta},\boldsymbol{\gamma})_{opt} = -1$—since $|c_0^{(r)}|^2 = 0$ means probability 1 of find marked elements—and otherwise the optimal angles are $\beta = \gamma = \pi$. To get the expectation value on the range $\rho \leq 0.25$, we compute $\phi_Y(\gamma)$ from the distribution $f_Y(x)$, given by

$$f_Y(x) = \begin{cases} \rho, & x = \rho - 1 \\ 1-\rho, & x = \rho, \end{cases} \tag{4.53}$$

since the mean of $f_X(x)$ is $-\rho$. Thus,

$$\begin{aligned} \phi_Y(\gamma) &= i\left(\sum_{x\in R_Y} f_Y(x)e^{-i\gamma x}\right)\left(\sum_{x\in R_Y} x f_Y(x)e^{i\gamma x}\right) \\ &= i[\rho e^{-i\gamma(\rho-1)} + (1-\rho)e^{-i\gamma\rho}][(\rho-1)\rho e^{i\gamma(\rho-1)} + \rho(1-\rho)e^{i\gamma\rho}] \\ &= i[e^{-i\gamma\rho}(\rho e^{i\gamma} + 1 - \rho)][\rho(1-\rho)e^{i\gamma\rho}(1-e^{-i\gamma})] \\ &= i\rho(1-\rho)((1-2\rho) + \rho(e^{i\gamma}+e^{-i\gamma}) - e^{-i\gamma}) \\ &= i\rho(1-\rho)((1-2\rho) + 2\rho\cos(\gamma) - \cos(\gamma) + i\sin(\gamma)) \\ &= i\rho(1-\rho)[(1-2\rho)(1-\cos(\gamma)) + i\sin(\gamma)]. \end{aligned} \tag{4.54}$$

Replacing $\gamma = \pi$, we get $\phi_Y(\pi) = 2i\rho(1-\rho)(1-2\rho)$. Thus, since $\phi_Y(\pi)$ is positive purely imaginary, $|\phi_Y(\pi)| = \mathrm{Im}\{\phi_Y(\pi)\}$ and by Eq. (4.15),

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = -\rho - 8\rho(1-\rho)(1-2\rho) = -\rho(16\rho^2 - 24p + 9) = -\rho(4\rho-3)^2, \tag{4.55}$$

and therefore,

$$E_1(\boldsymbol{\beta},\boldsymbol{\gamma})_{opt} = \begin{cases} -\rho(4\rho-3)^2, & \rho \leq 0.25 \\ -1, & \text{otherwise.} \end{cases} \tag{4.56}$$

The angle β obtained by using Eq. (4.16) is the expected optimal $\beta = \pi$.

Generalizing the result for arbitrary r is unfeasible through the analytical expression of Eq. (4.42). However, there is an alternative way based on the optimality of Grover's algorithm on unstructured search problem. To get that, firstly note that GM-QAOA applied to binary function is equivalent to the unstructured search problem with an arbitrary number of marked elements. This equivalence arises from the fact that there are r calls to an oracle for the binary function, and the objective of minimizing the expectation value aligns to maximize the probability of measuring a marked state. Not by coincidence, for $r = 1$ on $\rho \leq 0.25$ interval, the optimal angles $\beta = \gamma = \pi$ reduces a GM-QAOA round to a Grover's iteration. Indeed, $\rho = 0.25$ is the ratio in which Grover's algorithm with a single round reaches the probability 1 on measuring a marked element, so that if $\rho \leq 0.25$, Grover's operators are optimal, and if $\rho > 0.25$, the angles of Eq. (3.28) make the fine-tuning not to exceed the point of probability 1.

The aforementioned ratio that reaches probability 1, named here *threshold ratio* and denoted as $\rho_{Th}(r)$ for arbitrary r, is the point of $\pi/2$ radians angle of the geometric interpretation of Grover's algorithm, showed on Subsec. 3.2.1. The value of the threshold ratio, as discussed on Subsec. 3.2.3, is $\rho_{Th}(r) = \sin^2(\pi/(4r+2))$. Up to this point, for any number of iterations, Theorem 6 guarantees that Grover's algorithm gives the maximal average probability for measuring a marked state on the unstructured search problem. Note that, for instance, the variational approach done by Morales, Tlyachev, and Biamonte [75] performs slightly better than Grover's algorithm because the marked element ratio of the instances surpassed $\rho_{Th}(r)$. Using that result, we generalize GM-QAOA performance on binary function for an arbitrary number of layers with a constant time expression on Theorem 12. We denote by $P(\rho, r)$ the optimal probability of measuring a marked element of the binary function with GM-QAOA, which is the negative of the optimal expectation value.

Theorem 12 *For any number of rounds r and a binary function with ratio ρ, the optimal probability $P(\rho, r)$ of measuring a marked element with GM-QAOA is*

$$P(\rho, r) = \begin{cases} \sin^2\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right), & \rho \leq \rho_{Th}(r) \\ 1, & \textit{otherwise.} \end{cases} \tag{4.57}$$

Proof: The first interval is the probability of Eq. (3.8), which follows from Grover's optimality on average probability, applicable to GM-QAOA since the expectation value of QAOA with Grover mixer operator is invariant under any permutation of states—i.e., the positions of marked elements—and the capacity of GM-QAOA emulates Grover's algorithm. To establish the other interval, we split into $\rho \leq \rho_{Th}(r-1)$ and $\rho > \rho_{Th}(r-1)$ cases. To the first, we set $\beta_j = \gamma_j = \pi$ for all $j < r$ and the angles of Eq. (3.29) for the rth layer. Recall that by Golden et al. [26] analysis, $P(\rho, r) = 1$ is achieved for $\Delta > 0$. We need to prove that this inequality holds on the interval $\rho_{Th}(r) < \rho \leq \rho_{Th}(r-1)$. The interval is entirely contained on the proved interval of Eq. (4.57) for $P(\rho, r-1)$. Therefore, since $(1-\rho)M$ gives the number of non-marked elements, the probability of measuring non-marked elements can be written as $1 - P(\rho, r-1) = (1-\rho)M(c_{0,\pi}^{r-1})^2$. Replacing it in Eq. (3.30) gives

$$\Delta = \frac{1}{M}\left(4\rho - \frac{1 - P(\rho, r-1)}{1-\rho}\right) = \frac{1}{M}\left(\frac{P(\rho, r-1) - (1-2\rho)^2}{1-\rho}\right). \tag{4.58}$$

Since $P(\rho, r-1)$ is increasing on range $\rho_{Th}(r) < \rho \leq \rho_{Th}(r-1)$, Δ also does. Thus, to establish the positivity of Δ is enough to prove that $\Delta = 0$ at the point $\rho = \rho_{Th}(r)$. Applying trigonometric identities $\cos(x + \pi/2) = -\sin(x)$, $\sin(-x) = -\sin(x)$, $2\sin^2(x) = 1 - \cos(2x)$, and $2\cos^2(x) = 1 + \cos(2x)$ on the Eq. (4.58) at that point results

$$\begin{aligned}
\Delta &= \frac{1}{M(1-\rho)}\left(\sin^2\left(\frac{\pi(2r-1)}{4r+2}\right) - \left(1 - 2\sin^2\left(\frac{\pi}{4r+2}\right)\right)^2\right) \\
&= \frac{1}{M(1-\rho)}\left(\frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi(2r-1)}{4r+2}\right) - \cos^2\left(\frac{2\pi}{4r+2}\right)\right) \\
&= \frac{1}{2M(1-\rho)}\left(-\cos\left(\frac{2\pi(1-2r)}{4r+2}\right) - \cos\left(\frac{4\pi}{4r+2}\right)\right) \\
&= \frac{1}{2M(1-\rho)}\left(\sin\left(\frac{\pi(2r-3)}{4r+2}\right) - \sin\left(\frac{\pi(2r-3)}{4r+2}\right)\right) = 0,
\end{aligned} \tag{4.59}$$

as desired. To finish, in the case where $\rho > \rho_{Th}(r-1)$, there exists k such that $1 \leq k < r$ on which $\rho_{Th}(k-1) \geq \rho > \rho_{Th}(k)$ (since $\rho_{Th}(0) = 1$). Probability 1 can be reached with the earlier attribution on the kth first layers and $\beta_j = \gamma_j = 0$ to the remainder parameters to make the operators trivial.

□

Note that by proving the optimality of the choice of all the angles on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ being equal to π on $\rho \leq \rho_{Th}(r)$ interval, we prove that the efficient method of parameter finding of Golden et al. [26] indeed finds the optimal angles for a fixed t.

The probability $P(\rho, r)$ admits to be written as a polynomial function in terms of ρ on $\rho \leq \rho_{Th}(r)$ interval. To obtain it, we need the trigonometric identity $\cos(\arcsin(x)) = \sqrt{1-x^2}$ and the expansion of $\sin(nx)$ in terms of a summation of products of $\sin(x)$ and $\cos(x)$ given by, for an integer $n \geq 1$,

$$\sin(nx) = \sum_{k=0}^{\lfloor\frac{n-1}{2}\rfloor}(-1)^k\binom{n}{2k+1}\sin^{2k+1}(x)\cos^{n-2k-1}(x), \tag{4.60}$$

an identity closely related to the Chebyshev polynomials [124]. With them, we can finally get

$$\begin{aligned}
P(\rho, r) &= \left(\sum_{k=0}^{r}(-1)^k\binom{2r+1}{2k+1}\sin^{2k+1}(\arcsin(\sqrt{\rho}))\cos^{2(r-k)}(\arcsin(\sqrt{\rho}))\right)^2 \\
&= \rho\left(\sum_{k=0}^{r}(-1)^k\binom{2r+1}{2k+1}\rho^k(1-\rho)^{r-k}\right)^2.
\end{aligned} \tag{4.61}$$

The higher order of a term inside the parenthesis is r. Squaring and multiplying by ρ gives the maximum term of order $2r+1$. For $r = 1$, Eq. (4.61) gives the first case of Eq. (4.56). We show also $r = 2$, in which

$$P(\rho, 2) = \rho(16\rho^2 - 20\rho + 5)^2. \tag{4.62}$$

4.2 GM-Th-QAOA analysis

Due to the simplicity of the binary phase separation on GM-Th-QAOA compared to a phase separation codifying a general cost function on GM-QAOA, and the results of the binary function on GM-QAOA, proved on Subsec. 4.1.4, the analysis of GM-Th-QAOA is much simpler than of the GM-QAOA. That way, Theorem 13 provides a formula for the expectation value of GM-Th-QAOA based on the optimal probability $P(\rho, r)$ of measuring a marked element on the binary function on GM-QAOA, proved on Theorem 12, assuming as input the optimal angles already established. Furthermore, instead of using the characteristic function, as done in GM-QAOA, the main statistical quantity of the expression is the conditional expectation, expressed in terms of $G_Y(\cdot)$ and $F_Y(\cdot)$. For technical purposes, we allow for threshold value any $t \in \mathbb{R}$ and even the limit on $t \to -\infty$ and $t \to \infty$ cases.

Theorem 13 *For any number r of layers in GM-Th-QAOA with optimal angles, the expectation value is given by*

$$E_r(t) = \mu - G_Y(T)\frac{1 - P(\rho, r)/F_Y(T)}{1 - F_Y(T)}, \tag{4.63}$$

where $T = t - \mu$ and $P(\rho, r)$ is the optimal probability of measuring a marked solution in the binary function of ratio $\rho = F_Y(T)$ on GM-QAOA, given by Eq. (4.57). *For $F_Y(T) = 0$ and $F_Y(T) = 1$, we consider the respective limits on Eq.* (4.63).

Proof: From Eq. (3.27), we can computed the expectation value of GM-Th-QAOA as

$$E_r(t) = |c_1^{(r)}|^2 \sum_{k \in S: c(k) \leq t} c(k) + |c_0^{(r)}|^2 \sum_{k \in S: c(k) > t} c(k). \tag{4.64}$$

Similarly with GM-QAOA analysis, these summations can be performed equivalently using the pmf of X. For each possible cost, $x \in R_X$, we count the number of solutions k such that $c(k) = x$, i.e., $M f_X(x)$. Then,

$$E_r(t) = M|c_1^{(r)}|^2 \sum_{x \in R_X:\ x \leq t} x f_X(x) + M|c_0^{(r)}|^2 \sum_{x \in R_X:\ x > t} x f_X(x). \tag{4.65}$$

Let m be the number of states smaller or equal to t. We assume for now that $0 < m < M$. We can link our expression with the statistical quantities $F_X(t)$ and

$G_X(t)$. Using $F_X(t) = m/M$, $1 - F_X(t) = (M - m)/M$ and the definition of $G_X(t)$,

$$E_r(t) = \frac{G_X(t)}{F_X(t)} m|c_1^{(r)}|^2 + \frac{\mu - G_X(t)}{1 - F_X(t)} (M - m)|c_0^{(r)}|^2. \tag{4.66}$$

Note that the probability $P(\rho, r)$ of measuring a state smaller or equal to t is $m|c_1^{(r)}|^2$ and the probability $1 - P(\rho, r)$ for states above t is $(M - m)|c_0^{(r)}|^2$. Replacing it in Eq. (4.66),

$$E_r(t) = \frac{G_X(t)}{F_X(t)} P(\rho, r) + \frac{\mu - G_X(t)}{1 - F_X(t)} (1 - P(\rho, r)). \tag{4.67}$$

The random variable Y is introduced on our expression by using the properties $F_X(t) = F_Y(T)$ and $G_X(t) = \mu F_Y(T) + G_Y(T)$. Thus,

$$\begin{aligned} E_r(t) &= \mu + \frac{G_Y(T)}{F_Y(T)} P(\rho, r) - \frac{G_Y(T)}{1 - F_Y(T)} (1 - P(\rho, r)) \\ &= \mu - G_Y(T) \frac{1 - P(\rho, r)/F_Y(T)}{1 - F_Y(T)}. \end{aligned} \tag{4.68}$$

Now, note that both $F_Y(T) = 0$ $(m = 0)$ and $F_Y(T) = 1$ $(m = M)$ gets $E_r(t) \to \mu$. The first follows from $G_Y(T) = 0$ and $P(\rho, r)/F_Y(T) \to (2r + 1)^2$ (recall the maximum amplification of the low-convergence regime of Subsec. 3.2.2), and for the last, replacing $P(\rho, r) = 1$ on Eq. (4.63) gives $E_r(t) = \mu + G_Y(T)/F_Y(T)$ with the limit holding from $G_Y(T) \to \mu_Y = 0$. That is the desired value since in both cases, the final state of GM-Th-QAOA is a uniform superposition.

□

Some aspects and consequences of the above theorem are worth commenting on. Firstly, given $F_Y(T)$, $G_Y(T)$, and μ, we have a formula to compute $E_r(t)$ with complexity independent of r, as establishes Corollary 4. That allows us to analyze distributions with an arbitrary number of layers, far beyond the exponential complexity expression of GM-QAOA, and look at the asymptotic behavior on the number of layers.

Corollary 4 *Given a fixed threshold value t, the complexity of the expectation value expression of GM-Th-QAOA, given by Eq.* (4.63)*, has complexity independent of the number of layers.*

Secondly, by Eq. (4.67) and the definition of conditional expectation,

$$E_r(t) = \mathrm{E}[X|X \leq t] P(\rho, r) + \mathrm{E}[X|X > t](1 - P(\rho, r)). \tag{4.69}$$

The above equation gives an important intuition on the operation of GM-Th-QAOA. The expectation value $E_r(t)$ is a weighted sum by $P(\rho, r)$ of the expected value of the two sets split by the threshold value t.

Proceeding, the intuitive notion that $E_r(t) < \mu$ on $0 < F_Y(T) < 1$ follows since $P(\rho, r)/F_Y(T)$ is larger than 1, and from the fact, recurrent from here, that $G_Y(T)$ is negative because the mean of Y is zero and always there is a least one element positive and one negative on its support (since we eliminate the degenerate distribution). We also extend the definition of η (recall Subsec. 3.2.2) to $\eta = P(\rho, r)/F_Y(T)$ to consider also ratios above the threshold ratio $\rho_{Th}(r)$.

Now, if $F_Y(T) \geq \rho_{Th}(r)$, then $P(\rho, r) = 1$ and we have immediately the Corollary 5 using definitions on the conditional expectation.

Corollary 5 *For any number r of layers in GM-Th-QAOA with optimal angles, if $F_X(t) \geq \rho_{Th}(r)$, the expectation value is given by*

$$E_r(t) = \mu + \frac{G_Y(T)}{F_Y(T)} = \mu + E\left[Y | Y \leq T\right] = E\left[X | X \leq t\right]. \tag{4.70}$$

Moreover, we can write the expression of $E_r(t)$ in a polynomial form on $F_Y(T) \leq \rho_{Th}$ interval by using Eq. (4.61) for $P(\rho, r)$. Let

$$Q(\rho, r) = \sum_{k=0}^{r-1} (-1)^k \binom{2r+1}{2k+1} \rho^k (1-\rho)^{r-k-1}. \tag{4.71}$$

Using the factorization $1 - x^n = (1-x)(\sum_{k=0}^{r-1} x^k)$, we can prove that the polynomial is of order $2r - 1$ by eliminating the denominator $1 - F_Y(T)$ with

$$\begin{aligned} E_r(t) &= \mu - G_Y(T)\left(\frac{1 - F_Y(T)^{2r}}{1 - F_Y(T)} - \frac{P(\rho, r)/F_Y(T) - F_Y(T)^{2r}}{1 - F_Y(T)}\right) \\ &= \mu - G_Y(T)\left(\sum_{k=0}^{2r-1} F_Y(T)^k - \frac{\left((1 - F_Y(T))Q(\rho, r) + (-1)^r F_Y(T)^r\right)^2 - F_Y(T)^{2r}}{1 - F_Y(T)}\right) \\ &= \mu - G_Y(T)\left(\sum_{k=0}^{2r-1} F_Y(T)^k - (1 - F_Y(T))Q^2(\rho, r) - 2(-1)^r Q(\rho, r) F_Y(T)^r\right). \end{aligned} \tag{4.72}$$

Eq. (4.72) gives a polynomial expression on $F_Y(T)$ multiplying $G_Y(T)$ on the non-trivial terms. The case $r = 1$ is significant due to its simplicity and therefore is established in Corollary 6, which follows combining Eq. (4.72) and Corollary 5.

Corollary 6 *For a single layer in GM-Th-QAOA with optimal angles, the expectation value is given by*

$$E_1(t) = \begin{cases} \mu + 8G_Y(T)(1 - 2F_Y(T)), & F_Y(T) \leq 0.25 \\ \mu + \frac{G_Y(T)}{F_Y(T)}, & \textit{otherwise.} \end{cases} \tag{4.73}$$

With Eq. (4.73), we can optimize analytically for some particular distributions.

We show also $r = 2$, in which the polynomial expression has a cubic order. So,

$$E_2(t) = \mu + 8G_Y(Y)(3 - 22F_Y(T) + 48F_Y(T)^2 - 32F_Y(T)^3) \tag{4.74}$$

on $F_Y(T) \leq \rho_{Th}(2)$ range. For higher r, the polynomial expression becomes progressively more complicated than the trigonometric. However, the polynomial form cannot be discarded since someone might find a utility in an eventual analytical proof.

Subsequently, note that Corollary 5 induces a tight lower bound on GM-Th-QAOA performance (upper bound on $E_r(t)$ since we are minimizing it) and an upper bound on t_{opt}, given both by Corollary 7.

Corollary 7 *For any number r of layers in GM-Th-QAOA, the optimal threshold value is bounded by $t_{opt} \leq \tau$, where τ is the minimum t in which $P(\rho, r) = 1$, and we have a tight bound in the optimal expectation value given by*

$$E_r(t)_{opt} \leq E[X|X \leq \tau] \leq \tau. \tag{4.75}$$

Proof: First, note that there always exists a t for which $P(\rho, r) = 1$ because taking t as the maximum solution gives $F_X(t) = 1$. Then, by the definition of conditional expectation, $E[X|X \leq t_1] \leq E[X|X \leq t_2]$ for $t_1 < t_2$ and therefore the minimum t gives the best expectation value among the candidates of threshold in which $P(\rho, r) = 1$. To conclude the tightness of the bound, we consider the binary function with a parameter ρ such that $P(\rho, r) = 1$. In that case, $\rho = F_X(\tau)$ and then $E_r(t) = E[X|X \leq \tau] = \tau = -1$. Finally, if $t_{opt} \neq \tau$, $F_X(t_{opt})$ is smaller than $\rho_{Th}(r)$ and therefore $t_{opt} < \tau$. □

Another upper bound on t_{opt}, the intuitive notion that $t_{opt} \leq \mu$, is proven in Corollary 8.

Corollary 8 *For any number r of layers in GM-Th-QAOA, the optimal threshold value is bounded by $t_{opt} \leq \mu$.*

Proof: Suppose by contradiction that $t_{opt} > \mu$. Thus, $T_{opt} > 0$ and by definition of $G_Y(\cdot)$, $|G_Y(T_{opt})| \leq |G_Y(u)|$, where u is the first $u \in R_X$ such that $u \leq \mu$. Therefore, by Eq. (4.63), as $F_Y(T_{opt}) \geq F_Y(u)$, we just need to prove that the factor

$$\frac{P(\rho, r)/F_Y(T) - 1}{1 - F_Y(T)} \tag{4.76}$$

is strictly decreasing on $F_Y(T)$. Defining $R = 2r + 1$ and $u = R \arcsin\left(\sqrt{\rho}\right)$ in a same way as in Appendix A, we can rewrite Eq. (4.76) as

$$\frac{\sin^2(u)/\sin^2(u/R) - 1}{1 - \sin^2(u/R)} = \frac{4\sin^2(u)}{\sin^2(2u/R)} - \sec^2(u/R) \tag{4.77}$$

using the trigonometric identity $\sin(2x) = 2\sin(x)\cos(x)$. On $0 < u \leq \pi/2$, the first term is strictly decreasing by the same argument used in Appendix A on η case and is direct that $-\sec^2(u/R)$ also does. □

To finish, analogously to GM-QAOA, Corollary 9 neutralizes the impact of the standard deviation of X with the auxiliary Z. The corollary follows directly from the properties $F_Y(T) = F_Z(T/\sigma)$ and $G_Y(T) = \sigma G_Z(T/\sigma)$.

Corollary 9 *For any number r of layers in GM-Th-QAOA with optimal angles, the expectation value is given by*

$$E_r(t) = \mu - \sigma G_Z(T/\sigma)\frac{1 - P(\rho, r)/F_Z(T/\sigma)}{1 - F_Z(T/\sigma)}, \tag{4.78}$$

where $\rho = F_Z(T/\sigma)$.

As well as the GM-QAOA, the GM-Th-QAOA also depends only on the moments of order beyond expectation and variance. Thus, we denote $C_r(t)$, where $E_r(t) = \mu - C_r(t)\sigma$, and $C^{Th}(r)$ as the maximum $C_r(t)$ achieved by GM-Th-QAOA.

4.2.1 Threshold curve problem

Recall the conjecture present on Subsec. 3.8.2 that the curve of the expectation value versus the threshold value for angles obtained by the procedure of Golden et al. [26] decreases monotonically up to a valley value and then increases monotonically. Indeed, the method of Golden et al. [26], for fixed t, gives the optimal angles, as we establish on Subsec. 4.1.4. We call that curve by *threshold curve* and the conjecture by *threshold curve problem.* Since Theorem 13 gives a closed-form expression for the expectation value, we can directly tackle the threshold curve problem. To include the possibility of the threshold curve being constant for a consecutive pair of points, the considered behaviors in our proof are non-increasing and non-decreasing monotonicity instead of strictly decreasing and strictly increasing, respectively. We proved that the threshold curve must change its monotonicity only one time by establishing the derivative change of the sign one time. Using the step function form of $F_Y(T)$ and $G_Y(T)$, we can extend results about monotonicity for the original discrete random variable since it preserves the monotone behavior between any pair of consecutive points of the support. That was done in Theorem 14, proved in Appendix B.

Theorem 14 *For any number of layers in GM-Th-QAOA, the threshold curve is monotonically non-increasing up to a valley value and monotonically non-decreasing from there.*

4.2.2 Asymptotic tight bound on quantile

An alternative metric on the performance of GM-Th-QAOA is the quantity $F_X(E_r(t))$, which corresponds to the quantile in which the expectation value of GM-Th-QAOA is associated (obviously, we can use the quantile for GM-QAOA or any QAOA variant). That metric has as a strong point the possibility of comparing the obtained result with the spectrum of distribution itself. An immediate upper bound on $F_X(E_r(t))$ can be obtained by applying the cdf to both sides of the inequality in Corollary 7. Since cdf is a non-decreasing function, $F_X(E_r(t)_{opt}) \leq F_X(\tau)$. If we assume a continuous distribution, there is a t in which $F_X(t) = \rho_{Th}(r)$ for all r and then $F_X(\tau) = \rho_{Th}(r)$ for any r. That way, $F_X(E_r(t)_{opt})$ is bounded by $\rho_{Th}(r)$ and therefore

$$F_X(E_r(t)_{opt}) \leq \sin^2\left(\frac{\pi}{4r+2}\right) = \mathcal{O}\left(\frac{1}{r^2}\right). \tag{4.79}$$

The assumption X as a continuous random variable is convenient since discussing quantiles is more naturally suited for such distributions. We demonstrate in Theorem 15 that the asymptotic bound of Eq. (4.79) is tight. To do so, we rely on the supposition that R_X^{min} has a finite and non-zero value in pdf. That assumption is also quite reasonable since all target problems of QAOA have a finite optimal value, and the limits $f_X(R_X^{min}) \to 0$ or $f_X(R_X^{min}) \to \infty$ are just convenient mathematical abstractions in some situations.

Theorem 15 *For GM-Th-QAOA, if X is a continuous distribution and $f_X(R_X^{min}) = a$, where $0 < a < \infty$, then the quantile achieved by the optimal expectation value is asymptotically given by*

$$F_X(E_r(t)_{opt}) = \Theta\left(\frac{1}{r^2}\right). \tag{4.80}$$

Proof: The upper bound has already been established on Eq. (4.79). For the lower bound, let t be a fixed optimal threshold. We claim that $F_X(t) = F_Y(T) = \Theta(1/r^2)$. The bound $F_X(t) = \mathcal{O}(1/r^2)$ follows from $F_X(t) \leq \rho_{Th}(r)$ and to prove $F_X(t) = \Omega(1/r^2)$, consider the expectation value asymptotically on r knowing that $F_X(t) = \mathcal{O}(1/r^2)$, which gives, from Eq. (4.63),

$$E_r(t)_{opt} \to \mu + \mathrm{E}[Y|Y \leq T]P(\rho, r). \tag{4.81}$$

If $F_X(t) \notin \Omega(1/r^2)$,

$$P(\rho, r) = \sin^2\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right) \to \sin^2\left(2r\sqrt{\rho}\right) \to 0, \tag{4.82}$$

and as $\mathrm{E}[Y|Y \leq T]$ must be bounded by assumption (if $R_X^{min} \to -\infty$, we must have $a \to 0$), $\mathrm{E}[Y|Y \leq T]P(\rho, r) \to 0$ and then $E_r(t)_{opt} \to \mu$, which of course is not an optimal threshold, concluding by contradiction that $F_X(t) = \Omega(1/r^2)$.

Now, consider the bound

$$E_r(t)_{opt} \geq \mu + \mathrm{E}[Y|Y \leq T]P(\rho, r) \geq \mu + \mathrm{E}[Y|Y \leq T] = \mathrm{E}[X|X \leq t]. \tag{4.83}$$

The first inequality holds from $G_X(t) \leq 0$ on the first equality of Eq. (4.68) and the second by $P(\rho, r) \leq 1$. Applying cdf on both slides of Eq. (4.83) gives $F_X(E_r(t)_{opt}) \geq F_X(\mathrm{E}[X|X \leq t])$. Using the relation

$$0 < \lim_{n\to\infty} \frac{f(n)}{g(n)} < \infty \;\Rightarrow\; f(n) = \Theta(g(n)) \text{ as } n \to \infty, \tag{4.84}$$

since $F_X(t) = \Theta(1/r^2)$, by transitivity, we just need to prove that the limit

$$L = \lim_{t\to R_X^{min}} \frac{F_X(\mathrm{E}[X|X \leq t])}{F_X(t)} \tag{4.85}$$

is always non-zero finite. Note that denoting by $X_{\leq t}$ the random variable X given $X \leq t$, we have

$$\frac{F_X(\mathrm{E}[X|X \leq t])}{F_X(t)} = F_{X_{\leq t}}(\mathrm{E}[X|X \leq t]) = F_{X_{\leq t}}(\mathrm{E}[X_{\leq t}]), \tag{4.86}$$

since the pdf of $X_{\leq t}$ is $f_X(t)/F_X(t)$. Therefore, the statistical interpretation of the limit L is that it calculates the cdf of the expected value of $X_{\leq t}$ on $t \to R_X^{min}$. The limit is an indeterminate of $0/0$ type. The denominator follows from the definition of cdf, and for the numerator, using L'Hôpital's rule,

$$\lim_{t\to R_X^{min}} \mathrm{E}[X|X \leq t] = \lim_{t\to R_X^{min}} \frac{G_X(t)}{F_X(t)} = \lim_{t\to R_X^{min}} \frac{tf_X(t)}{f_X(t)} = R_X^{min}, \tag{4.87}$$

and then by the continuity of cdf,

$$\lim_{t\to R_X^{min}} F_X(\mathrm{E}[X|X \leq t]) = F_X\left(\lim_{t\to R_X^{min}} \mathrm{E}[X|X \leq t]\right) = F_X(R_X^{min}) = 0. \tag{4.88}$$

Applying L'Hôpital's rule in L gives

$$L = \lim_{t \to R_X^{min}} \frac{f_X\left(\frac{G_X(t)}{F_X(t)}\right) f_X(t) \frac{tF_X(t) - G_X(t)}{F_X(t)^2}}{f_X(t)} = a \lim_{t \to R_X^{min}} \frac{tF_X(t) - G_X(t)}{F_X(t)^2}, \tag{4.89}$$

where limit of $f_X(G_X(t)/F_X(t)) \to a$ follows by the continuity of pdf in the point R_X^{min}. We have another 0/0 indeterminate that follows immediately from the limit of Eq. (4.87). Therefore,

$$\begin{aligned} L = a \lim_{t \to R_X^{min}} \frac{tf_X(t) + F_X(t) - tf_X(t)}{2F_X(t) f_X(t)} &= \lim_{t \to R_X^{min}} \frac{F_X(t)}{2F_X(t)} = \lim_{t \to R_X^{min}} \frac{f_X(t)}{2f_X(t)} \\ &= \frac{1}{2}, \end{aligned} \tag{4.90}$$

as desired.

□

The theorem establishes a tight quadratic Grover-style speed-up of GM-Th-QAOA over classical brute force in the asymptotic limit, as it takes r rounds to attain an expectation value at a quantile of order $1/r^2$, in contrast to classical brute force, such as CRS, that with those number of round achieves a quantile of order of $1/r$. The result is expected since the optimal angles of GM-Th-QAOA reduce it to the execution of Grover's algorithm.

4.2.3 Upper bounds on the standard score for GM-Th-QAOA

Retaking the discussion of upper bounds on the standard score of Subsec. 4.1.3, now for GM-Th-QAOA, the problem the becomes feasible with our closed-form expression for $E_r(t)$. To begin with, similar to GM-QAOA, we provide the bound $C^{Th}(1) \leq 4$ with individual bounds. Specifically, applying $|G_Y(T)| \leq 0.5\sigma$ on Corollary 6 gives $C_1(t) \leq 4$ on $F_Y(T) \leq 0.25$ interval. It is unnecessary to check the other interval once by the definition of conditional expectation, setting $F_Y(T) = 0.25$, included on the limit of the other interval, gives the best bound on the range in which $P(\rho, r) = 1$ holds. To prove that the inequality holds, consider

$$\begin{aligned} \mathrm{E}[|Y|] &= \sum_{x \in R_Y} |x| f_Y(x) = -\sum_{x \in R_Y : x \leq 0} x f_Y(x) + \sum_{x \in R_Y : x > 0} x f_Y(x) \\ &= -G_Y(0) + (-G_Y(0)) = -2G_Y(0). \end{aligned} \tag{4.91}$$

The maximum of $|G_Y(\cdot)|$ is given when the argument is 0. Therefore $|G_Y(T)| \leq |G_Y(0)| = 0.5\,\mathrm{E}[|Y|] \leq 0.5\sigma$, as claimed.

For general r we can go far beyond the exponential bound of GM-QAOA. For that, using the bound of low-convergence regime $\eta \leq (2r+1)^2$ in addiction to

$|G_Y(T)| \leq 0.5\sigma$ on Theorem 13,

$$\begin{aligned} E_r(t) \geq \mu - \frac{(2r+1)^2-1}{2(1-F_Y(T))}\sigma \geq \mu - \frac{(2r+1)^2-1}{2(1-\rho_{Th}(r))}\sigma \\ = \mu - 2r(r+1)\sec^2\left(\frac{\pi}{4r+2}\right)\sigma, \end{aligned} \tag{4.92}$$

and we can conclude that

$$C^{Th}(r) \leq 2r(r+1)\sec^2\left(\frac{\pi}{4r+2}\right) = \mathcal{O}(r^2). \tag{4.93}$$

Indeed, the above bound is not tight, nor asymptotically. The tight upper bound is established through the assistance of Lemma 3, which claims that $C^{Th}(r)$ is attained by a particular family of distribution: the two-point distributions.

Lemma 3 *For any number r of layers in GM-Th-QAOA, the maximum standard score $C_r(t)$ achieved by GM-Th-QAOA, $C^{Th}(r)$, is hit by a two-point distribution.*

Proof: By Eq. (4.63), since $E_r(t) \leq \mu$ and $G_Y(T)$ is negative, for a fixed $F_Y(T)$, the maximum possible $C_r(t)$ is given when we maximize the ratio $|G_Y(T)|/\sigma$. Eliminating the trivial cases of $F_Y(T) = 0$ and $F_Y(T) = 1$ in which $C_r(t) = 0$, the key idea of the proof is that the distribution that maximizes that ratio is a two-point distribution for all the range $0 < F_Y(T) < 1$ and therefore $C^{Th}(r)$ necessarily be there.

To get it, we split the distribution into two parts by the threshold value defining $Y_{\leq T}$ as the random variable Y given $Y \leq T$ and $Y_{>T}$ as the random variable Y given $Y > T$. We also split the summation that computes σ^2 into two contributions

$$\sigma^2_{\leq T} = \sum_{x \in R_Y : x \leq T} x^2 f_Y(x), \ \sigma^2_{>T} = \sum_{x \in R_Y : x > T} x^2 f_Y(x), \tag{4.94}$$

where $\sigma = \sqrt{\sigma^2_{\leq T} + \sigma^2_{>T}}$. Thus, using the definitions of the conditional random variables,

$$\begin{aligned} \mathrm{E}[Y_{\leq T}] = \frac{G_Y(T)}{F_Y(T)}, \ \mathrm{E}[Y^2_{\leq T}] = \frac{\sigma^2_{\leq T}}{F_Y(T)}, \\ \mathrm{E}[Y_{>T}] = -\frac{G_Y(T)}{1-F_Y(T)}, \ \mathrm{E}[Y^2_{>T}] = \frac{\sigma^2_{>T}}{1-F_Y(T)}, \end{aligned} \tag{4.95}$$

to computed the bounds $\mathrm{E}[Y^2_{\leq T}] \geq \mathrm{E}[Y_{\leq T}]^2$ and $\mathrm{E}[Y^2_{<T}] \geq \mathrm{E}[Y_{<T}]^2$, we have

$$\frac{G_Y(T)^2}{F_Y(T)} \leq \sigma^2_{\leq T}, \ \frac{G_Y(T)^2}{1-F_Y(T)} \leq \sigma^2_{>T}. \tag{4.96}$$

Combining both bounds gives

$$\begin{aligned}&\frac{G_Y(T)^2}{F_Y(T)}+\frac{G_Y(T)^2}{1-F_Y(T)}\le\sigma_{\le T}^2+\sigma_{>T}^2\;\Rightarrow\;\frac{G_Y(T)^2}{F_Y(T)(1-F_Y(T))}\le\sigma^2\\&\Rightarrow\;\frac{|G_Y(T)|}{\sigma}\le\sqrt{F_Y(T)(1-F_Y(T))},\end{aligned}\tag{4.97}$$

and the equality is hit when the variance is 0, that is, if and only if both $Y_{\le T}$ and $Y_{>T}$ are single-point distributions, combining to get Y with two points. □

Since the random variable Z associated with any two-point distributions depends only on the ratio between the points, we can consider without loss of generality the binary function. Therefore, with Lemma 3, for a given r, $C^{Th}(r)$ can be founded by systematically varying the parameter ρ on binary function in the range $0<\rho<1$. The choice of threshold trivially is $t=-1$ for binary function, and in that way, $T_h(k)$ is precisely the original binary function. Therefore, since $E_r(t)=-P(\rho,r)$, from the definition of $C_r(t)$,

$$C_r(t)=\frac{P(\rho,r)+\mu}{\sigma}=\frac{P(\rho,r)-\rho}{\sqrt{\rho(1-\rho)}}.\tag{4.98}$$

Note that an alternative way to get Eq. (4.98) is to replace the bound of Eq. (4.97)—which is tight for binary function—on Eq. (4.63). For $r=1$, with the polynomial form of $P(\rho,1)$ on $\rho\le 0.25$ and 1 otherwise,

$$C_1(t)=\begin{cases}\frac{\rho(4\rho-3)^2-\rho}{\sqrt{\rho(1-\rho)}} & \rho\le 0.25\\ \frac{1-\rho}{\sqrt{\rho(1-\rho)}}, & \text{otherwise.}\end{cases}\tag{4.99}$$

For $\rho>0.25$, we can simplify to $C_1(t)=\sqrt{\frac{1-\rho}{\rho}}$ and since its derivative

$$\frac{dC_1(t)}{d\rho}=-\frac{1}{2\rho\sqrt{\rho(1-\rho)}}\tag{4.100}$$

is negative in the considered interval, the maximum is given on the limit at the point $\rho=0.25$. Therefore, we can ignore that interval since the point is included on $\rho\le 0.25$. Thus, for $\rho\le 0.25$, we manipulate $C_1(t)$ to

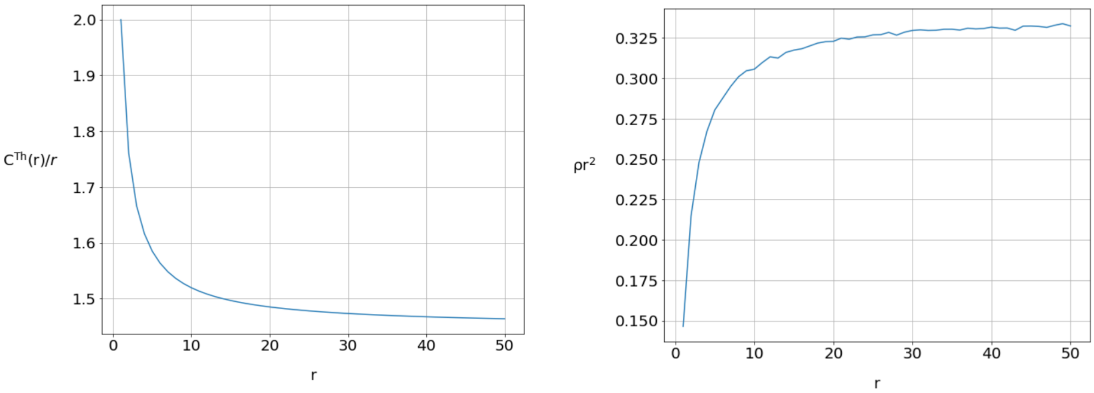
$$C_1(t)=\frac{\rho(4\rho-3)^2-\rho}{\sqrt{\rho(1-\rho)}}=\frac{8\rho(\rho-1)(2\rho-1)}{\sqrt{\rho(1-\rho)}}=8(1-2\rho)\sqrt{\rho(1-\rho)},\tag{4.101}$$

and take the derivative equalling zero giving

$$\frac{dC_1(t)}{d\rho}=-16\sqrt{\rho(1-\rho)}+\frac{4(1-2\rho)^2}{\sqrt{\rho(1-\rho)}}=\frac{4(8\rho^2-8\rho+1)}{\sqrt{\rho(1-\rho)}}=0.\tag{4.102}$$

Since the expression of the denominator has solutions $\rho = 0$ and $\rho = 1$, we only need to solve the quadratic equation on the numerator, which gives $\rho = \frac{\sqrt{2}\mp 1}{2\sqrt{2}}$. Since $\rho = \frac{\sqrt{2}+1}{2\sqrt{2}} > 0.25$, replacing the remainder solution $\rho = \frac{\sqrt{2}-1}{2\sqrt{2}}$ on Eq. (4.101) establishes the tight bound of $C^{Th}(1) = 2$.

We solve numerically for $r > 1$ using Nelder-Mead optimizer on the Python package *SciPy* [125]. The growth observed is linear in r, as shown by Fig. 4.2(a), which plots the ratio $C^{Th}(r)/r$ versus r up to 50 layers. The inclination of the linear curve converges to a value called κ. In fact, we prove in Theorem 16 that $C^{Th}(r) = \Theta(r)$ and the value of κ is a function of the solution of a transcendental equation, numerically evaluated to approximately 1.4482. In Fig. 4.2(b) we show, also up to $r = 50$, the curve of ρr^2 versus r, where ρ is the ratio what maximize $C_r(t)$. That quantity scales in the order $1/r^2$.



(a)

(b)

Figure 4.2: (a) The inclination of the curve $C^{Th}(r)$ versus r asymptotically converges to a certain value. (b) The ratio ρ that maximizes $C_r(t)$ scales with $1/r^2$, with a constant that asymptotically converges also to a certain value.

Theorem 16 *On the large limit of the number of layers r, the maximum standard score $C_r(t)$ achieved by GM-Th-QAOA is given by $C^{Th}(r) = \kappa r$, where $\kappa = 2\sin^2(x_1)/x_1$ for x_1 being the smallest positive solution of the equation $2x = \tan(x)$.*

Proof: By the same argument of $r = 1$ analysis, the optimal ρ satisfies $\rho \le \rho_{Th}(r)$. Then, in the large limit of r, Eq. (4.98) in that interval becomes $C_r(t) = \sin^2(2r\sqrt{\rho})/\sqrt{\rho}$. Taking the derivative of $C_r(t)$ equal to 0 gives

$$\begin{aligned} \frac{dC_r(t)}{d\rho} &= \frac{2r\sin(2r\sqrt{\rho})\cos(2r\sqrt{\rho}) - \frac{\sin^2(2r\sqrt{\rho})}{2\sqrt{\rho}}}{\rho} = 0 \\ \Rightarrow\ 2r\cos(2r\sqrt{\rho}) &= \frac{\sin(2r\sqrt{\rho})}{2\sqrt{\rho}} \ \Rightarrow\ 4r\sqrt{\rho} = \tan(2r\sqrt{\rho}). \end{aligned} \tag{4.103}$$

The substitution $x = 2r\sqrt{\rho}$ gives the trancendental equation $2x = \tan(x)$. The smallest positive solution $x \approx 1.1656$ gives $\rho \approx 0.3397/r^2$ (note that Fig. 4.2(b) gets close to that value of ρ), which is smaller than $\rho_{Th}(r) \to \pi^2/(16r^2)$. The next solution is $x \approx 4.604$ and does not obey $\rho \leq \rho_{Th}(r)$. Therefore, expressing $C_r(t)$ in terms of x_1 gives $C_r(t) = 2r\sin^2(x_1)/x_1$, and the theorem follows. □

Since the binary function is the same as GM-Th-QAOA in GM-QAOA, follow the lower bound $C^{GM}(r) \geq \kappa r$ on the limit of large r. In particular, for $r = 1$, combining with Theorem 11 gives $2 \leq C^{GM}(1) \leq \frac{8\sqrt{6}}{9}$.

Furthermore, the upper bound $C^{Th}(r)$ provides an explicit lower bound on the number of round r to reach a fixed approximation ratio λ, given by Corollary 10, that follows from the definitions of $C_r(t)$ and λ, respectively $E_r(t) = \mu - C_r(t)\sigma$ and $\lambda = E_r(t)/R_X^{min}$.

Corollary 10 *For any number r of layers in GM-Th-QAOA, provided that $R_X^{min} \neq 0$ and $|R_X^{min}| < \infty$,*

$$r \geq \frac{\mu - \lambda R_X^{min}}{(C^{Th}(r)/r)\sigma}. \tag{4.104}$$

In particular, on the large limit of r, $C^{Th}(r)/r = \kappa$.

To finish this subsection, we show another bound on the minimum rounds required to achieve an objective. Specifically, we get the minimum number of rounds for the algorithm finding the optimal with probability 1 (exact optimization). In that case, the optimal threshold must be $t_{opt} = R_X^{min}$ and we must satisfy $F_X(t_{opt}) = f_X(R_X^{min}) \geq \rho_{Th}(r)$. Therefore,

$$f_X(R_X^{min}) \geq \sin^2\left(\frac{\pi}{4r+2}\right) \Rightarrow r \geq \frac{1}{4}\left(\frac{\pi}{\arcsin\left(\sqrt{f_X(R_X^{min})}\right)} - 2\right), \tag{4.105}$$

and $f_X(R_X^{min}) \to 0$ gives

$$r \geq \frac{\pi}{4\sqrt{f_X(R_X^{min})}} = \Omega\left(\frac{1}{\sqrt{f_X(R_X^{min})}}\right), \tag{4.106}$$

a quadratic Grover-like speed-up.

4.2.4 Combining the bounds on the standard score and quantile

The explicitly tight bound on the standard score was built using different distributions for each r. In particular, the ratio ρ of the two-point distribution that hits $C^{Th}(r)$ changes with r. One can ask if a particular distribution gives an asymptotic optimal $C_r(t)$ of order $\Theta(r)$. If this were not the case, we would have the possibility of improving the bound of Corollary 10 for particular distribution on the asymptotic limit of r. However, we can get a family of distributions in which $C_r(t)$ scales arbitrarily close to $\Theta(r)$. The technique to obtain it consists of combining the bound of the quantile of Theorem 15 with the standard score $C_r(t)$.

To analyze the asymptotic behavior of $C_r(t)$ in terms of r, we must assume that $R_X^{min} \to -\infty$. However, since Theorem 15 has the supposition that $f_X(R_X^{min}) = a$, where $0 < a < \infty$, and $X \to -\infty$ gives $a \to 0$, is necessary the reasonable assumption that the limit L of Eq. (4.85) is non-zero finite. With the assumption on L, so that the result of Theorem 15 be applicable, analyzing the structure of its proof, remains to demonstrate that $F_X(t) = \Theta(1/r^2)$ on $R_X^{min} \to -\infty$ case, since the original one uses the premise of a finite R_X^{min}.

To get that, we assume initially a pdf with $f_X(-x) = \Omega(1/x^3)$. The minus sign in the argument is used to adapt to the standard asymptotic notation on $x \to \infty$. Recall the traditional definitions for asymptotic notation

$$\begin{aligned}
&f(n) = \mathcal{O}(g(n)) \text{ as } n \to \infty \;\Rightarrow\; \exists c > 0, \exists n_o, \forall n \geq n_o : f(n) \leq cg(n),\\
&f(n) = \Omega(g(n)) \text{ as } n \to \infty \;\Rightarrow\; \exists c > 0, \exists n_o, \forall n \geq n_o : f(n) \geq cg(n).
\end{aligned} \tag{4.107}$$

We fix the notation c and n_o to denote the constants on the asymptotic notation until the end of this subsection. With the second definition and the property $f_{-X}(x) = f_X(-x)$, we can bound

$$\begin{aligned}
\mathrm{E}[X^2] = \mathrm{E}[(-X)^2] &= \int_{-R_X^{max}}^{\infty} x^2 f_X(-x)\; dx\\
&= \int_{-R_X^{max}}^{n_o} x^2 f_X(-x)\; dx + \int_{n_o}^{\infty} x^2 f_X(-x)\; dx\\
&\geq \int_{-R_X^{max}}^{n_o} x^2 f_X(-x)\; dx + \int_{n_o}^{\infty} \frac{c}{x}\; dx \to \infty,
\end{aligned} \tag{4.108}$$

and therefore, X does not have a finite second moment, contradicting our assumption at the beginning of the chapter. Thus, we must have $f_X(-x) = \mathcal{O}(1/x^3)$. By the first definition of Eq. (4.107), choosing a x such that $x \geq n_o$,

$$\begin{aligned}
F_X(-x) &= \int_{-\infty}^{-x} f_X(k)\; dk = \int_{x}^{\infty} f_X(-k)\; dk \leq \int_{x}^{\infty} \frac{c}{k^3}\; dk = \frac{c}{2x^2}\\
&= \mathcal{O}\left(\frac{1}{x^2}\right).
\end{aligned} \tag{4.109}$$

We are interested in evaluating the asymptotic behavior of the quantile function, which, in our case, is the inverse cdf. Defining $H(x) = F_X(-x)$, by Eq. (4.107), $H(x) \leq h(x)$, where $h(x) = c/x^2$, for $x \geq n_o$. As $H(x)$ is a non-increasing function, $H^{-1}(x)$ also does and then $H^{-1}(H(x)) \geq H^{-1}(h(x))$. Since $H^{-1}(H(x)) = x =$

$h^{-1}(h(x))$, then $H^{-1}(h(x)) \leq h^{-1}(h(x))$. The inequality holds for $h(x) \leq h(n_o)$ (becomes from the original $x \geq n_o$). Setting $y = h(-x)$ and $y_o = h(-n_o)$, we have $H^{-1}(y) \leq h^{-1}(y)$ for $y \leq y_o$. Introducing the definition of asymptotic notation on the limit $x \to 0$,

$$\begin{aligned} f(n) = \mathcal{O}(g(n)) \text{ as } n \to 0 \;\Rightarrow\; \exists c > 0, \exists n_o, \forall n \leq n_o : f(n) \leq cg(n), \\ f(n) = \Omega(g(n)) \text{ as } n \to 0 \;\Rightarrow\; \exists c > 0, \exists n_o, \forall n \leq n_o : f(n) \geq cg(n), \end{aligned} \tag{4.110}$$

we see that $H^{-1}(y) = \mathcal{O}(h^{-1}(y)) = \mathcal{O}(1/\sqrt{y})$ as $y \to 0$. From the definition of $H(t)$, $H^{-1}(y) = -F_X^{-1}(y)$ and we conclude the asymptotic bound of $|F_X^{-1}(y)| = \mathcal{O}(1/\sqrt{y})$ as $y \to 0$ for the inverse cdf. As $F_X(\mathrm{E}[X|X \leq x])$ scales like $F_X(x)$ by the assumption on L,

$$\begin{aligned} |\mathrm{E}[X|X \leq x]| &= |F_X^{-1}(F_X(\mathrm{E}[X|X \leq x]))| \\ &= \mathcal{O}\left(\frac{1}{\sqrt{F_X(\mathrm{E}[X|X \leq x])}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{F_X(x)}}\right) \end{aligned} \tag{4.111}$$

as $F_X(x) \to 0$.

Let t be a fixed optimal angle for r rounds of GM-Th-QAOA. Since $F_X(t) \leq \rho_{Th}(r)$, from Eq. (4.57), $P(\rho, r)$ depends on $F_X(t)$ like $\Theta(F_X(t))$ as $F_X(t) \to 0$. By Eq. (4.81), $|E_r(t)_{opt}|$ is maximized by maximization the product of $|\mathrm{E}[Y|Y \leq T]|$ and $P(\rho, r)$. Note that if we decrease $F_X(t)$, $P(\rho, r)$ also decreases at the same time that $|\mathrm{E}[Y|Y \leq T]|$ increases. However, growth of $|\mathrm{E}[Y|Y \leq T]|$, bounded by Eq. (4.111), cannot compensate the decay of $P(\rho, r)$ and then $|E_r(t)_{opt}|$ is maximized assuming the slowest decay of $F_X(t)$. Therefore, since $F_X(t) \leq \rho_{Th}(r)$, $F_X(t) = \Theta(1/r^2)$ and we can extend the result of Theorem 15 to the Theorem 17.

Theorem 17 *For GM-Th-QAOA, if X is a continuous distribution and the limit L, given by Eq. (4.85), is non-zero finite, then the quantile achieved by the optimal expectation value is asymptotically given by*

$$F_X(E_r(t)_{opt}) = \Theta\left(\frac{1}{r^2}\right). \tag{4.112}$$

Establishes the applicable of the scale $1/r^2$ on quantile for our assumption, we can follow up with the built of a distribution in which $C_t(t)$ scales arbitrarily close to $\Theta(r)$. Then consider a $\epsilon > 0$ such that $f_Z(-x) = \Theta(1/x^{3+\epsilon})$. Repeating the previous argument we find $F_Z(-x) = \Theta(1/x^{2+\epsilon})$ and $|F_Z^{-1}(y)| = \Theta(1/\sqrt[2+\epsilon]{y})$ as $y \to 0$. So, for a fixed optimal threshold t, $F_Z(-C_r(t)) = F_X(E_r(t)) = \Theta(1/r^2)$ makes us conclude that $C_r(t) = \Theta(r^{2/(2+\epsilon)})$, in which $\epsilon \to 0$ gives exponent 1. An explicit distribution is

RPareto$(2+\epsilon, x_m)$ for $\epsilon > 0$, given by

$$f_X(x) = \frac{(\epsilon+2)x_m^{\epsilon+2}}{(-x)^{\epsilon+3}},\ x \in (-\infty, -x_m]. \tag{4.113}$$

The cdf and $G_X(\cdot)$ are computed with basic integration techniques as

$$F_X(x) = \frac{x_m^{\epsilon+2}}{(-x)^{\epsilon+2}},\ G_X(x) = -\frac{\epsilon+2}{\epsilon+1}\frac{x_m^{\epsilon+2}}{(-x)^{\epsilon+1}}, \tag{4.114}$$

and $\mathrm{E}[X|X \le x]$ is given by $\mathrm{E}[X|X \le x] = \frac{\epsilon+2}{\epsilon+1}x$. Thus, the limit L is

$$L = \lim_{x\to-\infty} \frac{(-x)^{2+\epsilon}}{\left(-\frac{\epsilon+2}{\epsilon+1}x\right)^{2+\epsilon}} = \left(\frac{\epsilon+2}{\epsilon+1}\right)^{-(2+\epsilon)} = \left(1+\frac{1}{1+\epsilon}\right)^{-(2+\epsilon)}. \tag{4.115}$$

That limit lies between 0.25 and $1/e$, with $L = 0.25$ in $\epsilon \to 0$ and resorting to the traditional definition of the Euler's number, $L = 1/e$ in $\epsilon \to \infty$.

In general, the optimal $C_r(t)$ asymptotically depends on r as $C_r(t) = \Theta(|F_Z^{-1}(1/r^2)|)$. If a distribution presents a cdf $F_Z(x)$ with exponential decay on $x \to -\infty$, the growth of $C_r(t)$ must be logarithm. It is the case of important distributions of literature, such as the normal, Laplace, reflected gamma, and reflected exponential distributions. Therefore, for an optimization problem with a probability distribution that exhibits a tendency of exponential decay, the number of rounds to achieve a fixed approximation ratio must be exponentially larger than the tight bound of Corollary 10 once

$$C_r(t) = \frac{\mu - \lambda R_X^{min}}{\sigma}. \tag{4.116}$$

Chapter 5

Numerical experiments

In this chapter, we provide numerical experiments computing the formulas of Theorem 10 and 13 for GM-QAOA and GM-Th-QAOA, respectively, with different probability distributions to emphasize important aspects of our analytical results. Except for the cases of a single layer for particular distributions in which it is possible to solve analytically and for discrete distributions in which we can use the method of Golden et al. [26] to find the optimal threshold value, we optimize the angles on GM-QAOA and the threshold value on GM-Th-QAOA by using the Nelder-Mead optimizer with the package *SciPy*. Also, in some situations, *SciPy* was used to compute statistical quantities. Due to the exponential complexity of the expression of the Theorem 10, we simulate GM-QAOA up to 8 layers. In contrast, for GM-Th-QAOA, since the expression has complexity independent of the number of layers, we simulate until 10^6 rounds on some occasions.

To the quadratic asymptotic speed-up of Theorem 17 be applicable, the limit L of Eq. (4.85) must be finite non-zero. That is the case for distributions studied in this chapter. Summarizing the results, for all considered probability distributions with $R_X^{min} \to -\infty$, except reflected Pareto distribution, the limit is $L = 1/e$. In most of them, such as Laplace, reflected exponential, and reflected Pareto (given on Eq. (4.115)) distributions, the limit can be evaluated analytically. The remainder cases can be verified on algebraic software, as emphasized in Sec. 1.2.

5.1 Normal distribution

We begin our numerical experiments with the normal distribution. The study of that distribution is justified by the ubiquity generated by the central limit theorem. In combinatorial optimization context, empirically has been observed that the solution space of the Capacitated Vehicle Routing and Portfolio Optimization seems to be normally distributed [23, 24]. It would not be surprising if other optimization problems were normally distributed. For this reason, we use the normal distribution as an example to illustrate many relevant issues.

The parameters u and s on distribution Normal(u, s^2) are location and scale parameters, respectively. Recall, in particular, they are the mean and standard deviation of the distribution, respectively. Therefore, the random variable Z is fixed, in such a way that, due to the Corollaries 3 and 9, we can assume without loss of generality $u = 0$ and $s = 1$ to the analysis.

5.1.1 Single layer

We consider initially $r = 1$. For GM-Th-QAOA, since the cdf of the normal distribution involves the error function [53], that only can be evaluated numerically, we optimize the threshold value numerically, resulting in $C_r(t)_{opt} \approx 1.346$ and $T_{opt}/\sigma = -0.8769$. Fig. 5.1 shows the threshold curve for $r = 1$ on the normal distribution.

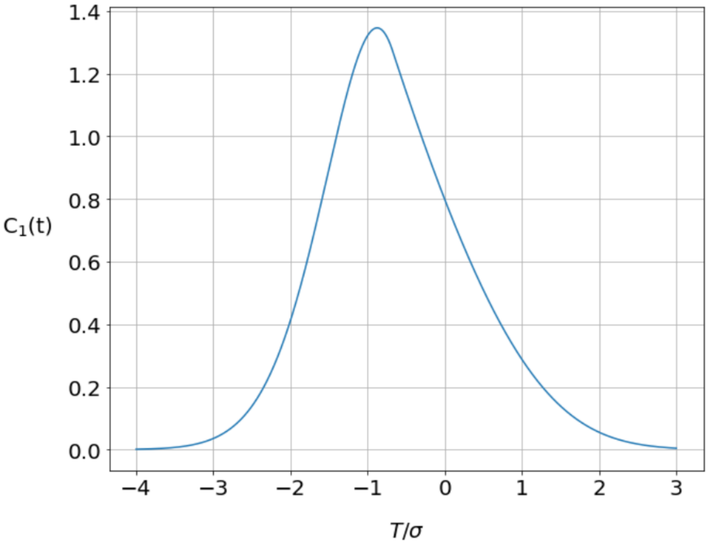


Figure 5.1: Threshold curve for $r = 1$ on normal distribution. Without loss of generality, we consider $C_r(t)$ versus T/σ as the threshold curve instead of the original $E_r(t)$ versus T of the Subsec. 4.2.1. T/σ means the normalized threshold for any choice of u and s. The resolution used is 2000 values of threshold, chosen uniformly between the interval of −4 and 3.

For GM-QAOA, on the other, we can optimize analytically. The characteristic function of the normal distribution is well-known, with $\varphi_Z(\omega) = e^{\frac{-\omega^2}{2}}$. This is the only case in this dissertation in which we demonstrate a statistical quantity of a particular distribution. Between existing methods to compute the integration, we choose the following: using the parity of $f_Z(x)$ we have

$$\varphi_Z(\omega) = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} e^{i\omega x}\, dx = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \cos(\omega x)\, dx. \tag{5.1}$$

Then, by the Leibniz rule for differentiation under integral sign,

$$\frac{d\varphi_Z(\omega)}{d\omega} = \int_{-\infty}^{\infty} \frac{d}{d\omega} \left(\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \cos(\omega x) \right) \, dx = \int_{-\infty}^{\infty} -x \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \sin(\omega x) \, dx. \tag{5.2}$$

Next, using integration by parts,

$$\frac{d\varphi_Z(\omega)}{d\omega} = \frac{\sin(\omega x)}{\sqrt{2\pi}} \left[e^{-\frac{x^2}{2}} \right]_{x=-\infty}^{\infty} - \omega \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \cos(\omega x) \, dx = -\omega \varphi_Z(\omega). \tag{5.3}$$

The solution of the resulting differential equation can be found, for instance, using the separation of variables method in such a way that $\varphi_Z(\omega) = Ke^{\frac{-\omega^2}{2}}$, where K is the constant to determine. To get that, we use as the initial condition the property $\varphi_Z(0) = 1$. Thus, $K = 1$ and $\varphi_Z(\omega) = e^{\frac{-\omega^2}{2}}$, as desired.

Taking the derivative, $\varphi'_Z(\omega) = -\omega e^{\frac{-\omega^2}{2}}$, and therefore

$$\phi_Z(\omega) = -\omega e^{-\omega^2}. \tag{5.4}$$

Since the normal distribution is a symmetric distribution, we can use Corollaries 1. Combine it with Corollary 3, and taking $\omega = \sigma\gamma$ we have $C_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 2\omega e^{-\omega^2}$. It is direct to check with the derivative $C_r(\boldsymbol{\beta}, \boldsymbol{\gamma})$ that

$$C_1(\boldsymbol{\beta}, \boldsymbol{\gamma})_{opt} = \sqrt{\frac{2}{e}} \approx 0.8578 \text{ with } (\beta, \sigma\gamma)_{opt} = \pm \left(-\frac{\pi}{2}, \frac{\sqrt{2}}{2} \right). \tag{5.5}$$

Fig. 5.2 shows the landscape of the normal distribution for a single round.

As a comparison, we can consider the algorithm CRS. As Bennett et al. [24] suggest, the equivalent computational effort to be used on CRS is $2r$, once the general procedure of Childs [109] to compile the phase separation operator consists of 2 oracle of the function $q(k)$. For $r = 1$, the expected value of the first order statistic of a random sample of size 2 on normal distribution is well-known and can be found, for instance, in Arnold, Balakrishnan, and Nagaraja [126] book. The value is $\mathrm{E}[X^{(1,2)}] = u - \frac{1}{\sqrt{\pi}} s$ and therefore, the standard score is $\frac{1}{\sqrt{\pi}}$.

Summarizing, for $r = 1$, the performance of CRS, GM-QAOA, GM-Th-QAOA on the normal distribution is in ascending order of standard scores with $\frac{1}{\pi}$, $\sqrt{\frac{2}{e}}$, and ≈ 1.346.

5.1.2 Discretization

Since the target of QAOA is COPs, which have a discrete domain by definition, it is natural to think of modeling the solution space as a discrete distribution. However, in some situations, it can be convenient to take the continuous case instead of

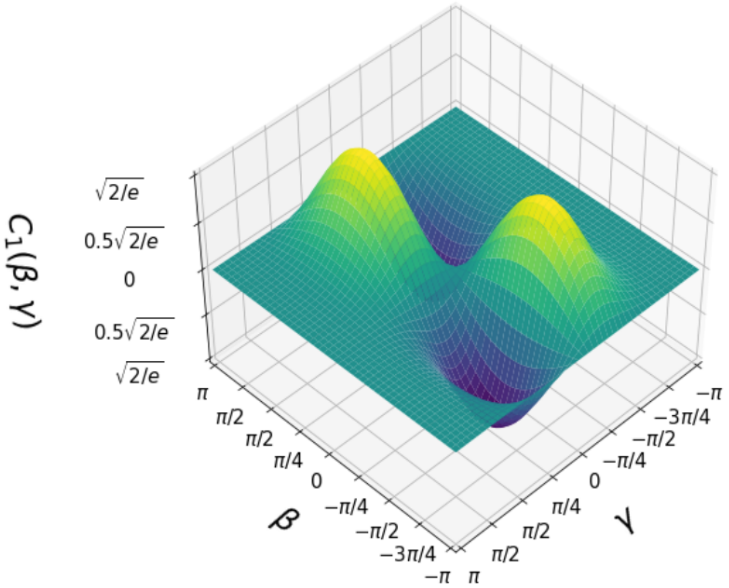


Figure 5.2: The landscape of GM-QAOA for $r = 1$ on the normal distribution. The range considered is $\gamma, \beta \in (\pi, \pi]$. We observe two symmetric global minima and two symmetric global maxima in the same way as discussed at the end of Subsec. 4.1.1.

dealing with the discrete solution space. In this subsection, we consider one of these situations, with the discretization of the normal distribution, showing that it can be quite complicated compared with the continuous ones and that the approximation of the continuous distributions can be quite accurate asymptotically. That raises doubts as to whether it is worth the effort to discretize.

The most natural approach to discretize the normal distribution is to discretize its support by considering only the integer values, which leads to the following summation (with $u = 0$ and $s = 1$) for the characteristic function,

$$\varphi_Z(\omega) = \sum_{x=-\infty}^{\infty} \frac{e^{\frac{x^2}{2}}}{\sqrt{2\pi}} e^{-i\omega x} = \frac{1}{\sqrt{2\pi}} \left(1 + 2 \sum_{x=1}^{\infty} e^{-\frac{x^2}{2}} cos(\omega x) \right) = \frac{\vartheta_3(\omega/2, e^{-\frac{1}{2}})}{\sqrt{2\pi}}, \tag{5.6}$$

where we use the parity of $\varphi_Z(\omega)$ and $\vartheta_3(z, q)$ is a Jacobi theta function defined by $\vartheta_3(z, q) = 1 + 2\sum_{x=1}^{\infty} q^{x^2} cos(2xz)$ [53], which is a non-elementary function, and therefore, not the best path to follow. Instead, we consider the discrete Gaussian kernel, DGK(s^2). From Lindeberg [52], DGK has mean 0 and standard deviation s, and therefore, it is appropriate to take the random variable Y. There is a closed-form expression for the characteristic function of DGK given by

$$\varphi_Y(\omega) = e^{s^2(\cos(\omega) - 1)}, \tag{5.7}$$

which has derivative $\varphi_Y'(\omega) = -s^2 \sin(\omega) e^{s^2(\cos(\omega)-1)}$. Then,

$$\phi_Y(\omega) = -s^2 \sin(\omega) e^{2s^2(\cos(\omega)-1)}. \tag{5.8}$$

Taking the derivative of the Eq. (5.8) and equalling to 0, we have

$$s^2 e^{2s^2(\cos(\omega)-1)}(2s^2 \sin^2(\omega) - \cos(\omega)) = 0. \tag{5.9}$$

The equation $s^2 e^{2s^2(\cos(\omega)-1)} = 0$ has no solution and then we can simplify the expression of Eq. (5.9) to

$$\begin{aligned}
&2s^2 \sin^2(\omega) - \cos(\omega) = 0 \Rightarrow\ \cos^2(\omega) + \frac{\cos(\omega)}{2s^2} = 1 \\
&\Rightarrow\ \left(\frac{1}{4s^2} + \cos(\omega)\right)^2 = 1 + \frac{1}{16s^4}\ \Rightarrow\ \frac{1}{4s^2} + \cos(\omega) = \pm\sqrt{1 + \frac{1}{16s^4}} \\
&\Rightarrow\ \omega = \pm\cos^{-1}\left(\pm\sqrt{1 + \frac{1}{16s^4}} - \frac{1}{4s^2}\right) = \pm\cos^{-1}\left(\frac{\sqrt{1+16s^4} - 1}{4s^2}\right),
\end{aligned} \tag{5.10}$$

where in the last equality, we eliminate half of the solutions because the argument of the arc cosine, in that case, is smaller than −1. Now, we consider the trigonometry identity $2\tan^{-1}(x) = \cos^{-1}\left(\frac{1-x^2}{1+x^2}\right)$. Set $x = \sqrt{\sqrt{1+16s^4} - 4s^2}$. Then,

$$\begin{aligned}
\frac{1-x^2}{1+x^2} &= \frac{1 - \sqrt{1+16s^4} + 4s^2}{1 + \sqrt{1+16s^4} - 4s^2} \\
&= \frac{(1 - \sqrt{1+16s^4} + 4s^2)(1 - \sqrt{1+16s^4} - 4s^2)}{(1 + \sqrt{1+16s^4} - 4s^2)(1 - \sqrt{1+16s^4} - 4s^2)} = \frac{\sqrt{1+16s^4} - 1}{4s^2},
\end{aligned} \tag{5.11}$$

and therefore we rewrite Eq. (5.10) as $\omega = \pm 2\tan^{-1}\left(\sqrt{\sqrt{1+16s^4} - 4s^2}\right)$. Our goal is to obtain the asymptotic behavior of the expression concerning the standard deviation s. To get it, we use the expansions by Taylor series $\sqrt{1+x} = 1 + x/2 + \mathcal{O}(x^2)$ as $x \to 0$ and $\tan^{-1}(x) = x + \mathcal{O}(x^3)$ as $x \to 0$,

$$\begin{aligned}
\omega &= \pm 2\tan^{-1}\left(2s\sqrt{\sqrt{1 + \frac{1}{16s^4}} - 1}\right) = \pm 2\tan^{-1}\left(2s\sqrt{\frac{1}{32s^4} + \mathcal{O}\left(\frac{1}{s^8}\right)}\right) \\
&= \pm 2\tan^{-1}\left(\frac{1}{2\sqrt{2}s} + \mathcal{O}\left(\frac{1}{s^3}\right)\right) = \pm\frac{\sqrt{2}}{2s} + \mathcal{O}\left(\frac{1}{s^3}\right).
\end{aligned} \tag{5.12}$$

The next step is to calculate the optimal value for the expectation value. Replacing the optimal $\omega = \pm\sqrt{2}/(2s) + \mathcal{O}(1/s^3)$ in Eq. (5.8) and using the Taylor series $\sin(x) = x + \mathcal{O}(x^3)$ as $x \to 0$ and $\cos(x) = 1 - x^2/2 + \mathcal{O}(x^4)$ as $x \to 0$, we get the optimal $\phi_Y(\omega)$

as

$$\phi_Y(\omega) = \left(\mp \frac{\sqrt{2}s}{2} + \mathcal{O}\left(\frac{1}{s}\right) \right) e^{-\frac{1}{2}+\mathcal{O}(1/s^2)} = \mp \frac{1}{2}\sqrt{\frac{2}{e}}s + \mathcal{O}\left(\frac{1}{s}\right). \tag{5.13}$$

Therefore,

$$C_1(\boldsymbol{\beta}, \boldsymbol{\gamma})_{opt} = \sqrt{\frac{2}{e}} + \mathcal{O}\left(\frac{1}{\sigma^2}\right) \text{ with } (\beta, \sigma\gamma)_{opt} = \pm\left(-\frac{\pi}{2}, \frac{\sqrt{2}}{2} + \mathcal{O}\left(\frac{1}{\sigma^2}\right)\right), \tag{5.14}$$

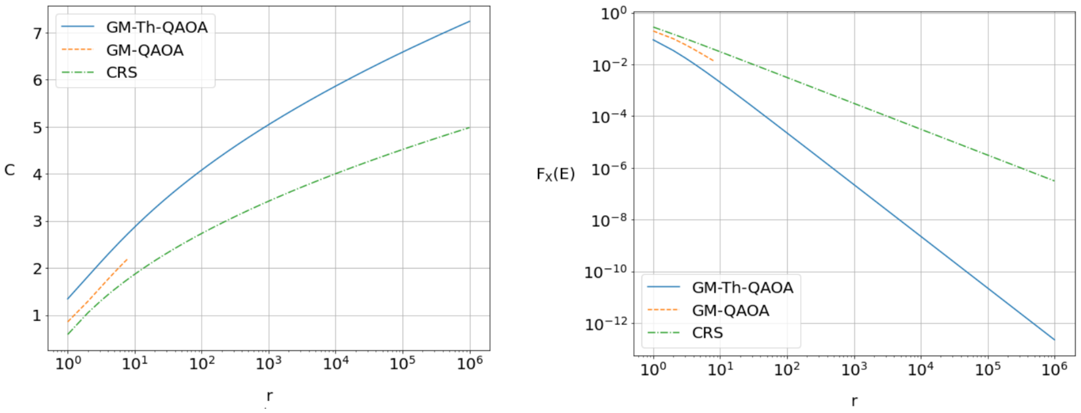
In the limit of large s, Eq. (5.14) is reduced to Eq. (5.5), which means the normal distribution is a quite accurate approximation for large instances (the usual target of QAOA) of combinatorial optimization problems in which the variance grows with the size of the input.

5.1.3 Scaling the number of layers

For $r > 1$, all algorithms are simulated numerically. To compare with CRS, we use Blom [127] asymptotic approximation for the expected value of the first order statistics, given by

$$\mathrm{E}[X^{(1,n)}] \approx u + F_N^{-1}\left(\frac{1-c}{n-2c+1}\right)s, \tag{5.15}$$

with $c = 0.375$ and where N is a random variable with standard normal distribution. To get the equivalent computational effort, we set $n = 2r$. Fig. 5.3 shows the simulation of distribution Normal(u, s^2) for GM-Th-QAOA and CRS up to large numbers of layers, and for GM-QAOA on the limit of simulation, considering in all cases the expected value and the cdf achieved by it. GM-Th-QAOA consistently overcomes GM-QAOA, as expected from the numerical results of the literature, and CRS, consistently with the quadratic gain. The asymptotic behavior of GM-Th-QAOA on $C_r(t)$ indicates a logarithmic growth, according to the expected from the exponential decay of the cdf on $x \to -\infty$, discussed on Subsec. 4.2.4. Furthermore, the asymptotic behavior of the cdf illustrates the quadratic gain of GM-Th-QAOA over classical brute force, given by Theorem 17, with the quantum algorithm scaling on a $1/r^2$ rate and the classical on $1/r$.

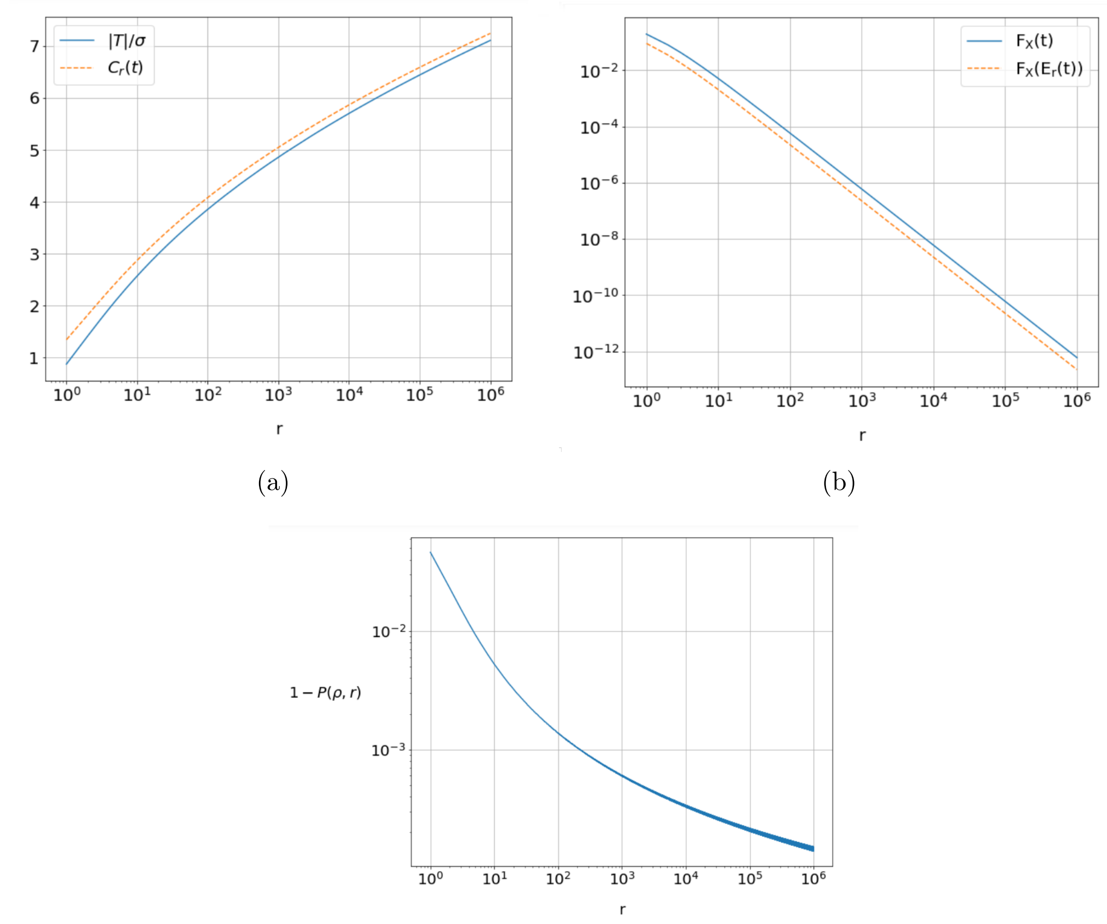


(a)

(b)

Figure 5.3: Simulation of distribution Normal(u, s^2) for GM-Th-QAOA and CRS up to 10^6 layers, and GM-QAOA up to 8 layers. (a) Standard score generically denoted by C versus r in a linear-log scale graphic. (b) Log-log graphic of the quantile achieved by the algorithms, generically denoted $F_X(\mathrm{E})$, as a function of r.

Fig. 5.4(a) illustrates the behavior of optimal threshold value, comparing it to the expectation value of GM-Th-QAOA. Although the threshold value has a larger value than the expectation value achieved by GM-Th-QAOA, both have the same scale $1/r^2$ in terms of quantile, as observed in Fig. 5.4(b), which is expected from the asymptotic behavior $F_X(t) = \Theta(1/r^2)$ obtained on the proof of Theorem 17. Fig. 5.4(c) shows for that optimal threshold value how the probability $P(\rho, r)$ scales with r. The graphic indicates that the algorithm acts closer to probability 1 as the number of layers increases.



(a)

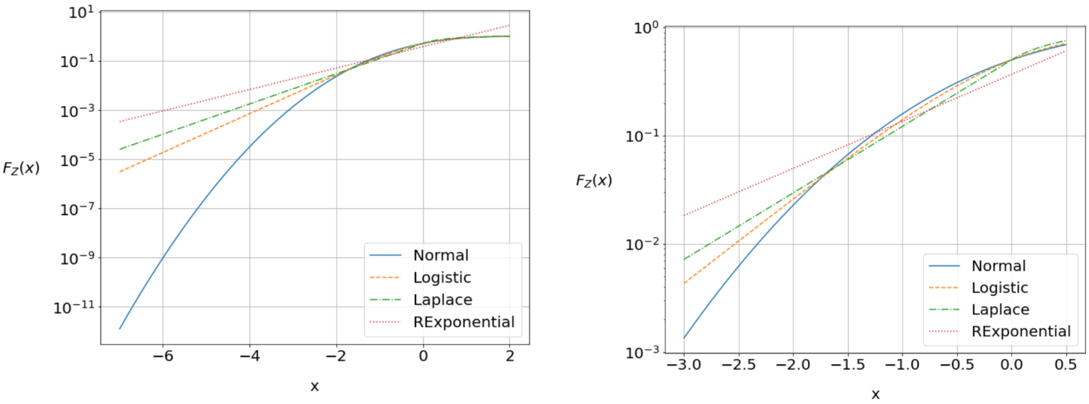
(b)

(c)

Figure 5.4: Optimal threshold on the simulation of distribution Normal(u, s^2) for GM-Th-QAOA up to 10^6 layers. (a) Linear-log graphic of $|T|/\sigma$ versus r, compared with $C_r(t)$. The absolute value of T is taken to compare directly with $C_r(t)$. (b) Log-log graphic of $F_X(t)$ versus r, compared with $F_X(E_r(t))$. (c) For that optimal threshold value, the figure shows the probability $P(\rho, r)$ on a graphic with a log-log scale for $1 - P(\rho, r)$ versus r.

5.2 Outliers effect

Recall the discussion on Subsec. 2.4.3 about kurtosis and outliers. The kurtosis is a measure connected with the propensity to produce outliers, which, in turn, are related in distributions of $R_X^{min} \to -\infty$, with the asymptotic decay of the cdf. As discussed on Subsec. 4.2.4, since the quantile evolves asymptotically with a fixed quadratic speed-up on GM-Th-QAOA, the standard score is determined by the decay of the cdf, with $C_r(t) = \Theta(|F_Z^{-1}(1/r^2)|)$. To illustrate that issue, we deal with the distributions Normal(u, s^2), Logistic(u, s), and Laplace(u, b), illustrated in Fig. 2.3, in addition to the reflected exponential distribution RExponential(l), which has standard probability distribution $f_Z(x) = e^{x-1}$ for any l—that is, a fixed random variable Z, such as the three other distributions. Fig. 5.5 shows that the cdf decay of reflected exponential distribution is slower than all three considered distributions, not by chance having the highest kurtosis value between them with $\mathrm{Kurt}[X] = 9$. Among the remainder distributions, recall that the ascending order of slower decay is normal, logistic, and Laplace. Therefore, the expected ascending order on the asymptotic values of $C_r(t)$ among the four considered distributions is normal, logistic, Laplace, and reflected exponential distributions.



(a) (b)

Figure 5.5: Decay of the cdf of the standard reflected exponential distribution, compared to the standard normal, standard logistic, and standard Laplace distributions. The scale of both graphics is log-linear. (a) Larger range of $x \in [-7, 2]$. (b) Zoom on the range $x \in [-3, 0.5]$.

We begin with GM-QAOA, a good starting point, although only scaled to something similar to ten rounds and has not been proven to have quadratic speed-up. Firstly, be worth mentioning that for $r = 1$, the Laplace distribution is one of the distributions that can be solved analytically. Analogously to the normal distribution, one can get

$$C_1(\boldsymbol{\beta},\boldsymbol{\gamma})_{opt} = \frac{25}{108}\sqrt{10} \approx 0.7230 \text{ with } (\beta,\sigma\gamma)_{opt} = \pm\left(-\frac{\pi}{2}, \sqrt{\frac{2}{5}}\right). \tag{5.16}$$

Retaking the discussion, Fig. 5.6 shows the performance of these four distributions up to 8 layers for GM-QAOA by showing $C_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ and $F_X(C_r(\boldsymbol{\beta},\boldsymbol{\gamma}))$. Among the distributions normal, logistic, and Laplace, as $F_X(C_r(\boldsymbol{\beta},\boldsymbol{\gamma}))$ evolves more or less similarly, the decays of Fig. 5.5 explain the standard score, starting with the ascending order of that quantity being normal, logistic, and Laplace, and ending reverting to Laplace, logistic, and normal. The overtaking points more or less coincide on both graphs of Fig. 5.5 and Fig. 5.6(a). On the other hand, the reflected exponential has a performance harder to interpret due to its asymmetry. For instance, the quantile of the mean is $1/e$, smaller than the 0.5th quantile of symmetric distributions, which means that the reflected exponential distribution starts with an "advantage" on small rounds of QAOA. Added to that, we have the slowest decay between all considered distributions which contributes to the standard score far superior to others. Fig. 5.7 plots the analogous graphics of Fig. 5.6 but for GM-Th-QAOA in a simulation up to 10^6 rounds. The expected result of the ascending order of the values of $C_r(t)$ for normal, logistic, Laplace, and reflected exponential distributions is obtained, with the cdf scaling as $1/r^2$ on larger rounds with an almost imperceptible difference.

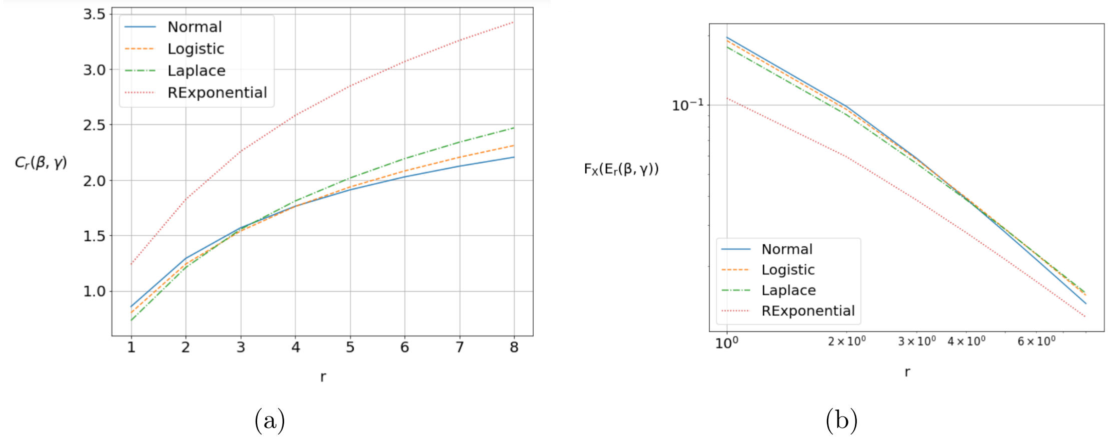


Figure 5.6: Simulation of normal, logistic, Laplace, and reflected exponential distributions for GM-QAOA up to 8 layers. (a) $C_r(\boldsymbol{\beta},\boldsymbol{\gamma})$ versus r. (b) $F_X(E_r(\boldsymbol{\beta},\boldsymbol{\gamma}))$ versus r in log-log scale.

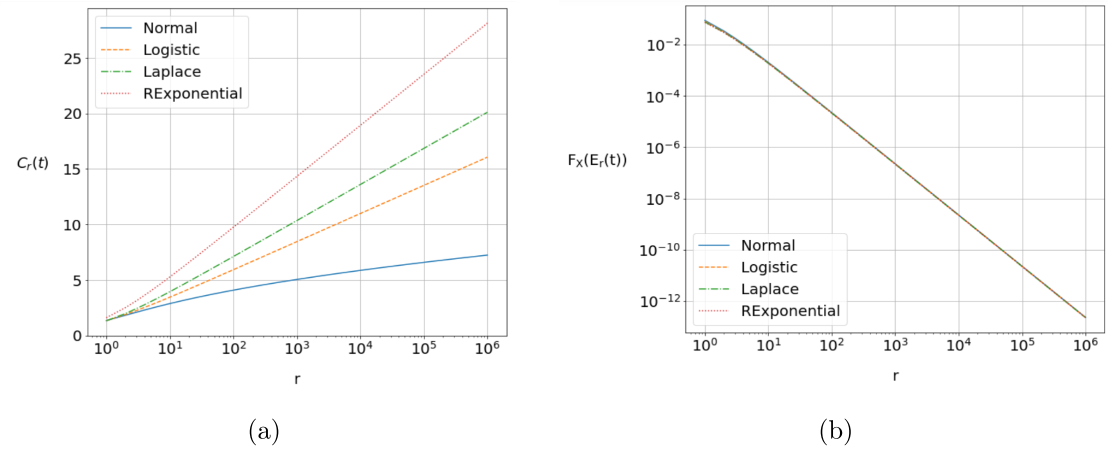


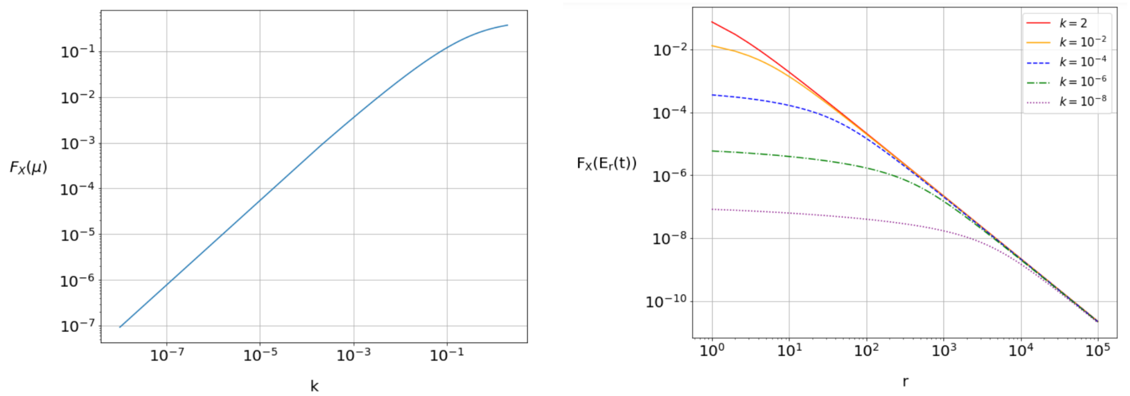
Figure 5.7: Simulation of normal, logistic, Laplace, and reflected exponential distributions for GM-Th-QAOA up to 10^6 layers. (a) $C_r(t)$ versus r on linear-log scale. (b) $F_X(E_r(t))$ versus r in log-log scale.

These results emphasize the importance of statistical moments of order from the third onwards on the performance of GM-QAOA and GM-Th-QAOA. In GM-Th-QAOA, the correlation with kurtosis is evident due to the fixed asymptotic quadratic speed-up. In GM-QAOA, until someone formally decides whether that variant also exhibits quadratic speed-up behavior, the relationship of the standard score with the moments is less known, although Eq. (4.48) guarantees that it exists.

To finish, although we have omitted, the analogous conclusions of the graphics of Fig. 5.4 can be obtained with logistic, Laplace, and reflected exponential distributions.

5.3 Asymptotic scale of the quadratic speed-up

To emphasize the asymptotic aspect of Theorem 17, we consider the reflected gamma distribution $\mathrm{RGamma}(a, b)$. For that, we simulate the distribution for values of a and b that progressively make it left-skewer—recall the discussion of Subsec. 2.4.3 and particularly on Fig. 2.2. That way, the quantile of the expected value $F_X(\mu)$ decreases so that QAOA already begins with a low quantile, as shown in Fig. 5.8(a), and evolves slowly in the first rounds. However, as Fig. 5.8(b) shows, given a sufficient number of rounds, the asymptotic scale of $F_X(E_r(t)) = \Theta(1/r^2)$ appears, since the cdf of the expected value of the distribution X given $X \leq t$ approaches the limit L.



(a)

(b)

Figure 5.8: Distribution $\mathrm{RGamma}(a, b)$ with $b = 1/2$ and $a = k/2$ for values of $k = 2$ and then decreasing with powers of 10 such that $k = 10^{-j}$ for $j = 2, 4, 6, 8$. Note that if k were a positive integer, we would have the chi-squared distribution of k degrees of freedom and that the case of $a = 2$ and $b = 1/2$ reduce the reflected distribution gamma to the reflected exponential distribution with $l = 1/2$. (a) $F_X(\mu)$ versus k on log-log scale from 10^{-8} to 2 with 10000 values of resolution. (b) Log-log graphic of $F_X(E_r(t))$ versus r up to 10^5 rounds.

Furthermore, although we have omitted, one can observe logarithm scales on $C_r(t)$ by plotting RGamma(a,b) on sufficient numbers of layers for the values of a and b considered in Fig. 5.8.

5.4 Near optimal asymptotic standard score

In Subsec. 4.2.4, we conclude that distributions that scales like reflected Pareto distribution RPareto$(2+\epsilon, x_m)$ with small ϵ, have asymptotic standard score close to the bound $C^{Th}(r) = \Theta(r)$ of Theorem 16. We numerically illustrate it in this section, specifically for the distribution RPareto$(2+\epsilon, x_m)$. To a desired $0 < j < 1$ such that $C_r(t) = \Theta(r^j)$, we can choose the parameter $\epsilon = 2(1-j)/j$. The parameter x_m is a scale parameter and, therefore, irrelevant to the analysis. Fig. 5.9 shows the simulation of RPareto$(2+\epsilon, x_m)$ on GM-Th-QAOA up to 10^5 rounds for the values $j = 0.1, 0.3, 0.5, 0.7, 0.9, 0.99$. Fitting all curves [128] with a power-law (using *SciPy*), we found the exponents $0.99, 0.9, 0.7703, 0.5023, 0.3136, 0.1570$ for the respective values of j in descending order. Although the behavior is more precise with the theoretical results on higher values of j, the confluence is just a matter of simulating sufficient numbers of layers. For instance, for $j = 0.1$, fitting on $r = 1$ up to $r = x$ for the range $x = 10, 10^2, 10^3, 10^4, 10^5$ gives the progressive improvement of respectively $0.5087, 0.3222, 0.2301, 0.1834, 0.1570$ on the coefficients.

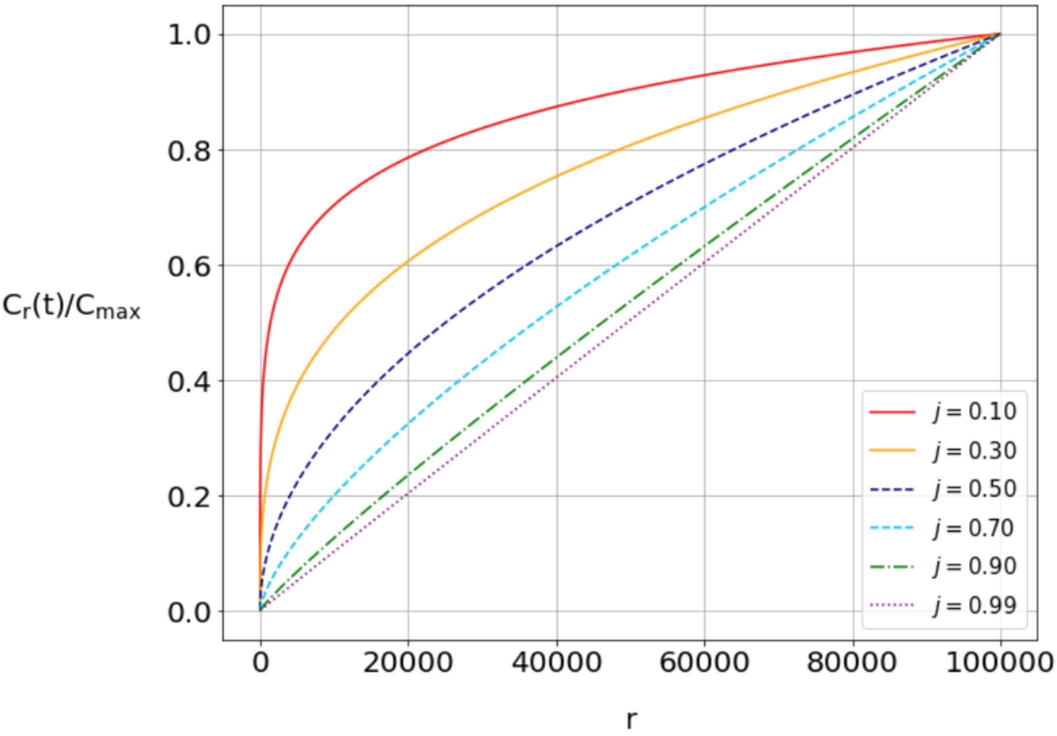


Figure 5.9: Standard score achieved by GM-Th-QAOA up to 10^5 rounds for Pareto(ϵ, x_m) with different values of j. For viewing purposes, we normalize $C_r(t)$ by C_{max}, where C_{max} is the value of $C_{10^5}(t)$.

Of course, $F_X(E_r(t))$ scales $1/r^2$ for the instances of Fig. 5.9, although it does not add to the content of the work to explicitly show the graphics.

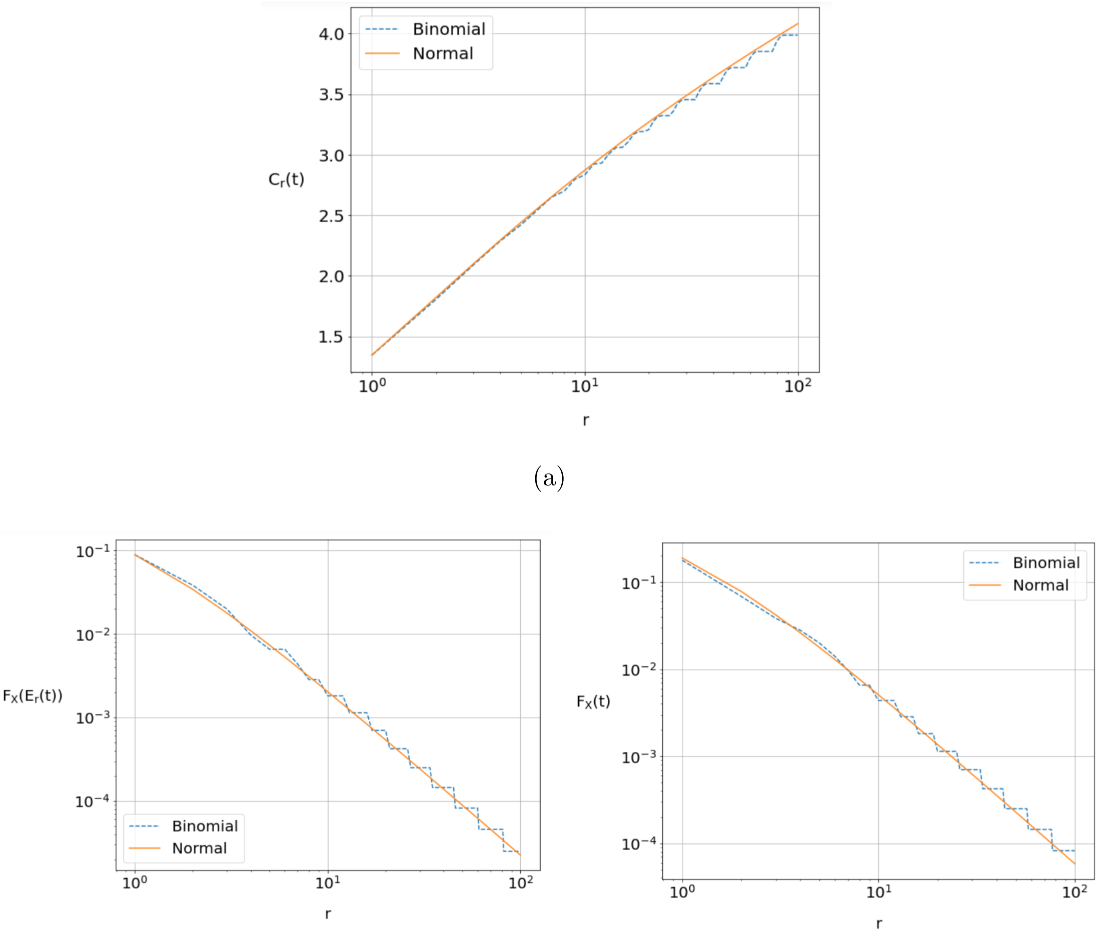
5.5 Discrete distributions

Until now, except for DGK on a single round for GM-QAOA, we only consider continuous distributions throughout this chapter. In this section, we address the differences between discrete and continuous distributions, directly comparing the binomial and discrete uniform distributions with normal and continuous uniform distributions, respectively. The similarity with the discrete uniform and continuous uniform distributions is immediate, and since Binomial(n, p) is the sum of n independent Bernoulli random variables with probability p, by the CLT, it approaches normal distribution on $n \to \infty$.

In GM-QAOA, since the statistical quantity is the characteristic function, both discrete and continuous distribution have a similar nature from the algorithm's point of view. On the other hand, for GM-Th-QAOA, a continuous spectrum for the threshold value changes the dynamics of the algorithm compared with a discrete set of choices of candidates of the threshold.

5.5.1 Binomial distribution

We choose the values of $n = 200$ and $p = 0.5$ for Binomial(n, p). The value of $p = 0.5$ is to get a symmetric distribution, and the value $n = 200$ is not a much larger so that the differences can be perceivable. Fig. 5.10 plots $C_r(t)$, $F_X(E_r(t))$, and $F_X(t)$ versus r for both binomial and normal distributions. As expected from the CLT, all scales similarly. However, note that from a certain r on the binomial distribution, $C_r(t)$, $F_X(E_r(t))$, and $F_X(E_r(t))$ do not grow for every increase in r, keeping stagnant for some rounds. The cases of $F_X(E_r(t))$ and $F_X(t)$ can be partially explained by the definition of the cdf on points outside the support R_X, but the complete picture is explained in Fig. 5.11, which shows the optimal threshold and its associated probability $P(\rho, r)$, both in a function of r.

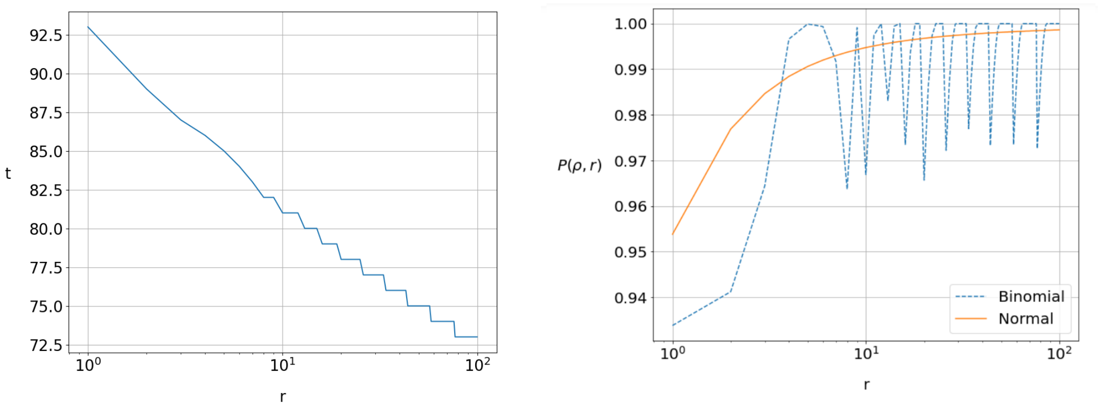


(a)

(b)

(c)

Figure 5.10: Simulation of Binomial(n, p) with $n = 200$ and $p = 0.5$ for GM-Th-QAOA up to 100 rounds, compared with the distribution Normal(u, s^2). (a) $C_r(t)$ versus r on the linear-log scale and (b) $F_X(E_r(t))$ versus r on log-log scale. (b) $F_X(t)$ versus r on log-log scale.

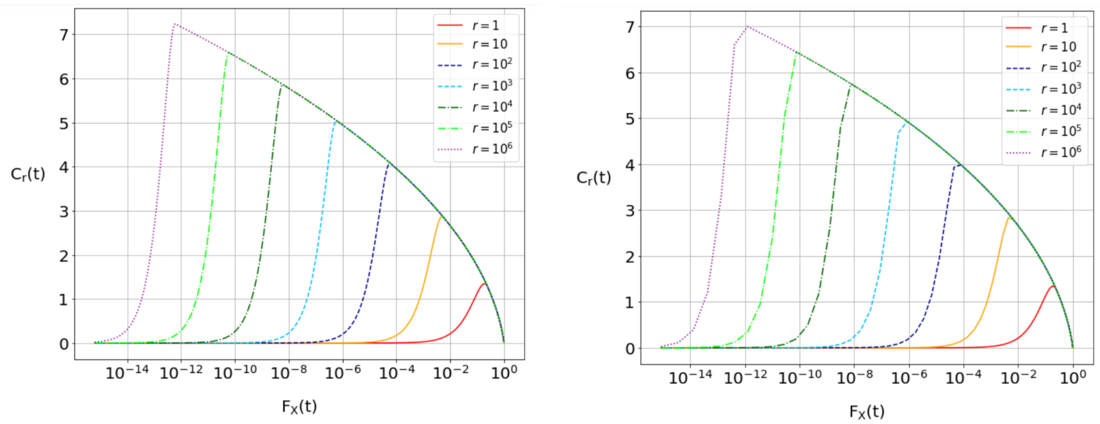


(a)

(b)

Figure 5.11: (a) The optimal threshold of GM-Th-QAOA and its probability (b) $P(\rho, r)$ (we also show $P(\rho, r)$ of the normal distribution as background) versus r in linear-log scale for the distribution Binomial(n, p) with $n = 200$ and $p = 0.5$ up to 100 rounds. By plot (a), the threshold value starts to stagnate for some rounds after a certain point. For a given value of optimal threshold t, evolving the number of rounds, the probability $P(\rho, r)$ increases until eventually arriving at the maximum value of 1, as observed in the plot of (b). From there, the only way to improve the performance of GM-Th-QAOA is by changing the threshold to the next value, $t - 1$. However, we may need more than one round for the change to be advantageous, and thus, the algorithm stagnates in that interval. Upon reaching $t - 1$, probability returns to below 1, and the process repeats, which explains the behavior of Fig. 5.10. Indeed, we can observe that the points with probability 1 of the plot (b) match the stagnation points of Fig. 5.10.

Fig. 5.12 shows the threshold curve of binomial and normal distributions for different values of r. Again, without loss of generality, we consider a different metric as the threshold curve, being now $C_r(t)$ versus $F_X(t)$. Both curves are similar and illustrate the result of Theorem 14, monotonically increasing $C_r(t)$ to up the optimal point and then monotonically decreasing.



(a) (b)

Figure 5.12: Threshold curve of distributions (a) Normal(u, s^2) and (b) Binomial(n, p) ($n = 200$ and $p = 0.5$) with $C_r(t)$ versus $F_X(t)$ on a linear-log scale. The resolution considered on the continuous distribution was of 2000 values for the threshold. For viewing purposes, we show only the values of r in terms of powers of 10 from 1 up to 10^6. The envelope that unites the curves is the interval of $P(\rho, r) = 1$.

5.5.2 Uniform distributions

We begin with the continuous uniform distribution, a distribution with fixed random variable Z. That is a distribution that, for GM-Th-QAOA, we can analytically optimize the threshold value for $r = 1$ by using the polynomial expression of Corollary 6. As the distribution is continuous, we can ignore the range $F_Y(T) > \rho_{Th}(r)$ by the analog reason as discussed on Subsec. 4.2.3. Combining Corollary 6 with Corollary 9, since $F_Z(x) = \frac{1}{6}(3 + \sqrt{3}x)$ and $G_Z(x) = \frac{x^2-3}{4\sqrt{3}}$, we have

$$C_1(t) = -8\frac{x^2 - 3}{4\sqrt{3}}\left(1 - \frac{1}{3}(3 + \sqrt{3}x)\right) = \frac{2}{3}x^3 - 2x. \tag{5.17}$$

with $x = T/\sigma$. It is direct to check with the derivative that

$$C_1(t)_{opt} = \frac{4}{3} \text{ with } T_{opt}/\sigma = -1. \tag{5.18}$$

Since $R_Z^{min} = -\sqrt{3}$, the maximum achievable by the standard score is $\sqrt{3}$. Fig. 5.13 shows the standard score obtained on the simulation of continuous uniform distribution in the algorithms GM-QAOA and GM-Th-QAOA up to 8 rounds. Again, GM-Th-QAOA consistently overcomes GM-QAOA.

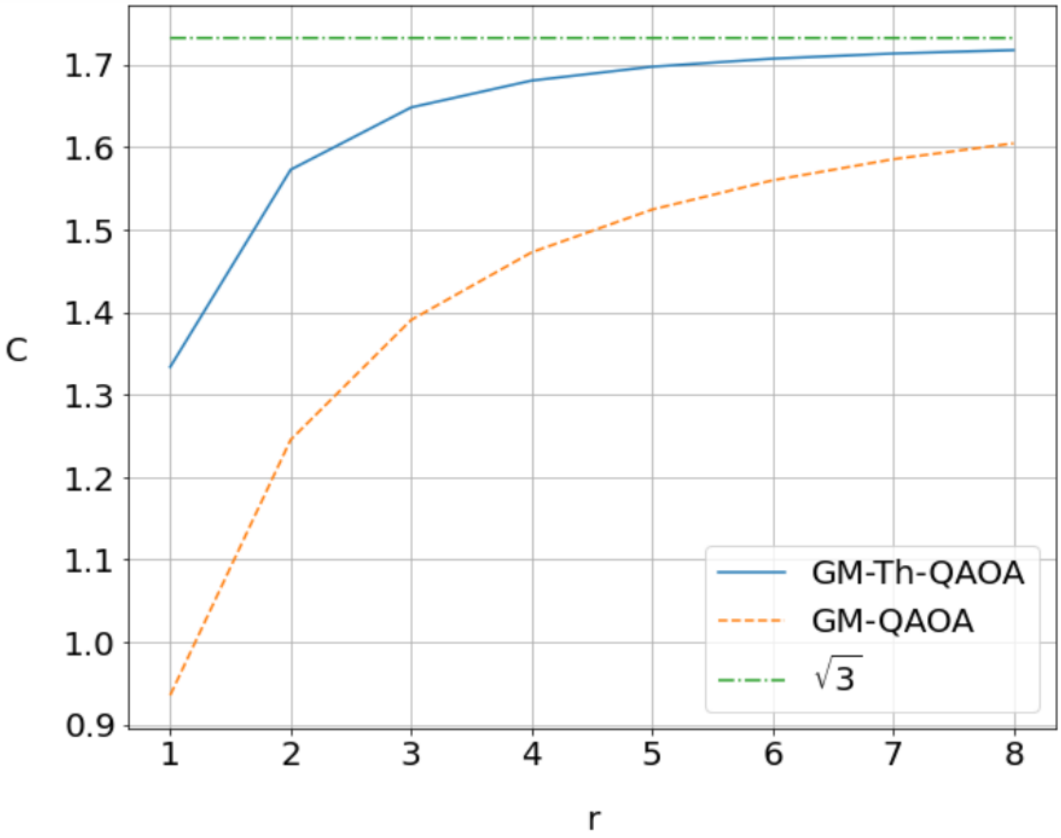


Figure 5.13: The standard score achieved by GM-QAOA and GM-Th-QAOA, generically denoted by C, versus r, up to 8 rounds for the distribution CUniform(a, b). We indicate the maximum achievable standard score $\sqrt{3}$ with a line segment.

For a finite number of rounds, we cannot achieve $C = \sqrt{3}$ on GM-Th-QAOA and GM-QAOA since on the continuous distribution approximation, the "number of states" to amplify is uncountable. On the other hand, for discrete distribution, such as the case of DUniform(a, b), we can get the optimal solution with probability 1 since it involves a countable number of states. The discrete uniform distribution does not have a fixed random variable Z. Thus, taking $n > 0$ for the distribution DUniform$(-n, n)$, since the mean is 0 and the standard deviation is $\sqrt{\frac{n(n+1)}{3}}$, we have

$$R_Z^{min} = -\sqrt{\frac{n}{n+1}}\sqrt{3}. \tag{5.19}$$

Note that if $n \to \infty$, then $R_Z^{min} \to -\sqrt{3}$. For a finite n, GM-Th-QAOA can find the optimal standard score with a finite number of layers—follows since $f_Z(R_Z^{min}) \geq \rho_{Th}(r)$ for some finite r. In this sense, Fig. 5.14 compares $C_r(t)$ for the continuous uniform distribution and for the discrete uniform distribution with different values of n. We see that while the continuous distribution evolves progressively without reaching its optimal, each discrete distribution gets the optimal on the minimum value given by Eq. (5.19) for some number of rounds.

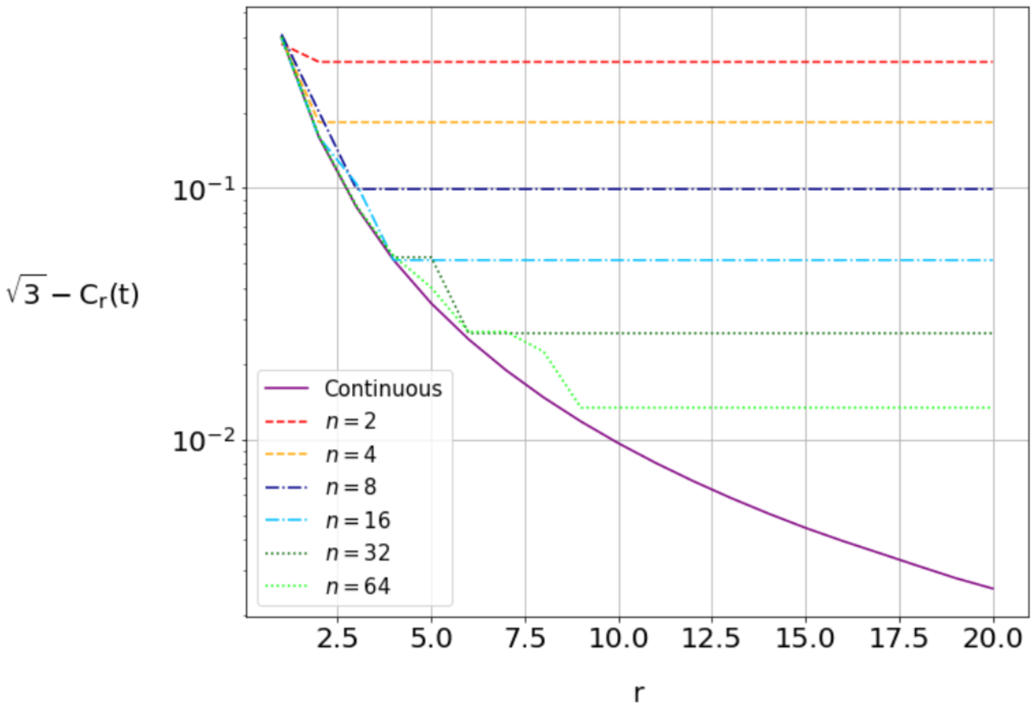


Figure 5.14: Standard score achieved by GM-Th-QAOA up to 20 layers for the distributions CUniform(a, b) and DUniform$(-n, n)$, with n increasing with powers of 2 such as 2^j for $j = 1, 2, 3, 4, 5, 6$. The graphic is a log-linear of $\sqrt{3} - C_r(t)$ versus r.

Chapter 6

Bounds on Grover-based Quantum Alternating Operator Ansatz

In Subsec. 4.2.2, we prove an asymptotic tight bound that implies GM-Th-QAOA has a quadratic speed up over the classical brute force approach. That raises the question of whether that bound is general for any variant of QAOA with Grover mixer, whether GM-QAOA or any potential new variation that can emerge, an issue related to the discussion of Sub. 3.11. To answer that question, we extend the definition of QAOA with the Grover mixer for the aforementioned Grover-based QAOA, encompassing a phase separation operator that encodes any real-valued function compiled from the cost function. Using that generalization, we develop a technique to get a general upper bound that consists of getting the maximum amplification of the probability over any set of degenerate states. With that upper bound established, we explicitly construct the minimum expectation value within that constrained framework.

6.1 The Grover-based Quantum Alternating Operator Ansatz

From Def. 17 of QAOA, we define Grover-based QAOA as follows.

Definition 20 (Grover-based QAOA) *Grover-based QAOA is the particular case of QAOA on Def. 17 in which the mixer Hamiltonian is the Grover mixer, given by Eq.* (3.23)*; the initial state is* $|s\rangle$*, given by Eq.* (3.24)*; and the goal is to minimize expectation value* $\langle\psi^{(r)}|H_C|\psi^{(r)}\rangle$*, particularly denoted* E_r*.*

Note that the Theorem 7 is applicable to that framework. For GM-QAOA, the function $q(k)$ is precisely the cost function, and for GM-Th-QAOA, $q(k) = T_h(k)$. The statistical analysis, introduced in Chapter 4, can be applied to Grover-based.

For that, we introduce the subscript on random variables to differ between the distributions of the functions $c(k)$ and $q(k)$. Specifically, for the originals X, Y, and Z we respectively denoted X_c, Y_c, and Z_c for $c(k)$, and X_q, Y_q, and Z_q for $q(k)$. The random variable X_q can be expressed as a mapping from X_c such that $X_q = q(X_c)$. For instance, in GM-Th-QAOA, we have

$$X_q = q(X_c) = \begin{cases} -1, & X_c \leq t \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

Applying the analogous analysis of GM-QAOA, the characteristic function factors now refer to the random variable X_q, such that $\varphi_X(\gamma)$ becomes $\varphi_{X_q}(\gamma)$, while the derivative of the characteristic function is changed from $\varphi'_X(\gamma)$ to $\Psi_X(\gamma)$, where

$$\begin{aligned}\Psi_X(\gamma) &= i\langle s|H_C U_P^\dagger(\gamma)|s\rangle = i\frac{1}{M}\sum_{k\in S} c(k)e^{i\gamma q(k)} \\ &= i\sum_{x\in R_{X_c}} x f_{X_c}(x)e^{i\gamma q(x)} = i\sum_{y\in R_{X_q}} e^{i\gamma y}\sum_{x\in R_{X_c}:q(x)=y} x f_{X_c}(x).\end{aligned} \tag{6.2}$$

We do not find a statistical interpretation of the quantity $\Psi_X(\gamma)$. The symbols μ and σ continue to denote the mean and standard deviation associated with the cost function. In particular, $\mu = -i\Psi_X(0)$, and to change from Y to Z, we have, in a similar way to the derivative of the characteristic function,

$$\begin{aligned}\Psi_Y(\gamma) &= i\sum_{x\in R_{Y_c}} x f_{Z_c}(x/\sigma)e^{i\gamma q_Y(x)} = i\sigma\sum_{x\in R_{Z_c}} x f_{Z_c}(x)e^{i\gamma q_Y(\sigma x)} \\ &= i\sigma\sum_{x\in R_{Z_c}} x f_{Z_c}(x)e^{i\gamma\sigma q_Z(x)} = \sigma\Psi_Z(\sigma\gamma),\end{aligned} \tag{6.3}$$

where $q_Y(x)$ and $q_Z(x)$ are the analogs of $q(x)$ to Y and Z, respectively. Using directly the aforementioned changes, we generalize Theorem 10 and Corollary 3 to Theorem 18, which provides three expectation value expressions, one as a function of each random variable X_c/X_q, Y_c/Y_q, and Z_c/Z_q, given from Eq. (4.40), (4.42), and (4.47), respectively. For the standard score, we denote C_r, where $E_r = \mu - C_r\sigma$, and $C(r)$ denotes the maximum C_r achieved by Grover-based QAOA.

Theorem 18 *For any number r of layers in Grover-based QAOA, the expectation value is given*

$$\begin{aligned}E_r = -i\sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} &\left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right]\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right] \\ &\left[\prod_{\mathcal{P}\in\mathcal{P}_L}\varphi_{X_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_R}\varphi^*_{X_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\Psi_X\left(\sum_{j\in\mathcal{P}_0}\gamma_j\right)\right],\end{aligned} \tag{6.4}$$

$$
\begin{aligned}
E_r = \mu + 2\,\mathrm{Im}\Bigg\{ & \sum_{\boldsymbol{x}^{(L)}<\boldsymbol{x}^{(R)}:L^{max}<R^{max}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right]\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right] \\
& \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_{Y_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi^*_{Y_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\Psi_Y\left(\sum_{j\in\mathcal{P}_0}\gamma_j\right)\right]\Bigg\},
\end{aligned}
\tag{6.5}
$$

$$
\begin{aligned}
E_r = \mu + 2\sigma\,\mathrm{Im}\Bigg\{ & \sum_{\boldsymbol{x}^{(L)}<\boldsymbol{x}^{(R)}:L^{max}<R^{max}} \left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right]\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right] \\
& \left[\prod_{\mathcal{P}\in\mathcal{P}_L} \varphi_{Z_q}\left(\sigma\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_R} \varphi^*_{Z_q}\left(\sigma\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\Psi_Z\left(\sigma\sum_{j\in\mathcal{P}_0}\gamma_j\right)\right]\Bigg\}.
\end{aligned}
\tag{6.6}
$$

6.2 General bounds

To bound the maximum amplification of the probability over any set of degenerate states, we define S_T as a set of elements on the spectrum of the Hamiltonian H_Q with some fixed cost x_o. Suppose that H_Q is built from an arbitrary problem Hamiltonian H_C. For a given r, our goal is to maximize the ratio between the probability of measuring a state on S_T before and after the application of the QAOA operators optimizing the choices of the ratio $|S_T|/M$ and the probability distribution $f_{X_q}(x)$. The only restriction on the choice of the distribution $f_{X_q}(x)$ is sign the probability $|S_T|/M$ on value x_o. To get it, consider taking the expectation value on the final state of Grover-based QAOA of a third Hamiltonian H_{max} that encodes x_o to $|S_T|$ elements and 0 to the remainders, with ratio $\rho = |S_T|/M$. The probability of measuring an element of S_T on the initial state is ρ, while after the application of QAOA operators is $E_r^{max}(S_T, f_{X_q})/x_o$, where $E_r^{max}(S_T, f_{X_q})$ denotes the expectation value of that configuration. We want to maximize the ratio between then, named $\eta_r(S_T, f_{X_q})$, and use it bound to explicitly build the minimum expectation value on an arbitrary instance of some Grover-based QAOA by sequentially maximally amplifying the states in ascending order of cost until the sum of probabilities reaches 1. As the amplitudes of degenerate states are equal, the amplification is in ascending order of the support of X_c.

For $r = 1$, we get the maximum amplification analytically. The mean and $\Psi_X(\gamma)$ are with respect to H_{max}, giving $\mu = x_o\rho$ and

$$
\Psi_X(\gamma) = i\langle s|H_{max}U_P^\dagger(\gamma)|s\rangle = ix_o\rho e^{i\gamma x_o}, \tag{6.7}
$$

Consequently, by Eq. (6.4),

$$\begin{aligned} &E_1^{max}(S_T, f_{X_q}) = x_o\rho + |B(\beta)|^2|\varphi_{X_q}(\gamma)|^2 x_o\rho + 2\,\mathrm{Im}\{ix_o\rho e^{i\gamma x_o}B(\beta)\varphi^*_{X_q}(\gamma)\} \\ &\Rightarrow\ \eta_1(S_T, f_{X_q}) = 1 + |B(\beta)|^2|\varphi_{X_q}(\gamma)|^2 + 2\,\mathrm{Re}\{e^{i\gamma x_o}B(\beta)\varphi^*_{X_q}(\gamma)\}. \end{aligned} \tag{6.8}$$

Since $|B(\beta)| \leq 2$ and $|\varphi_{X_q}(\gamma)| \leq 1$, then $\eta_1(S_T, f_{X_q}) \leq 9$. That value is saturated if $\rho \to 0$ and the remainder probability of $f_{X_q}(x)$ is completed on value 0, i.e., if $f_{X_q}(x)$ represents a binary function up to a scale change of ratio ρ, in which $\beta = \pi$ and $\gamma = \pi/x_o$ is optimal. In particular, it can be seen with $\varphi_{X_q}(\pi/x_o) \to 1$ and $B(\pi) = -2$.

Note that the maximum amplification is $(2r + 1)^2$, the exact amplification of Grover's algorithm on the low-convergence regime. One can ask if the maximum amplification is on the low-convergence regime for any r. Recall from Sec. 3.10 that there is numerical evidence for that from Bennett and Wang [23] in the context of QWOA on the complete graph. Unfortunately, applying the individual bounds of $|B(\beta)|$ and $|\varphi_{X_q}(\gamma)|$ on general r expression gives an amplification of exponential order, which does not help us. To be more precise, following steps similar to those of $r = 1$ case, we get the general expression

$$\begin{aligned} \eta_r(S_T, f_{X_q}) = \sum_{\boldsymbol{x}^{(L)},\boldsymbol{x}^{(R)}} \left[\exp\left(ix_o\sum_{j\in\mathcal{P}_0}\gamma_j\right)\right]\left[\prod_{j:x_j^{(L)}=1} B^*(\beta_j)\right]\left[\prod_{j:x_j^{(R)}=1} B(\beta_j)\right] \\ \left[\prod_{\mathcal{P}\in\mathcal{P}_L}\varphi_{X_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right]\left[\prod_{\mathcal{P}\in\mathcal{P}_R}\varphi^*_{X_q}\left(\sum_{j\in\mathcal{P}}\gamma_j\right)\right], \end{aligned} \tag{6.9}$$

and apply the individual bounds given, from the analysis of Subsec. 4.1.3, $\eta_r(S_T, f_{X_q}) \leq 9^r$. Moreover, like in the standard score bound of GM-QAOA, direct analytical treatment is unfeasible, necessitating indirect methods. Specifically, we demonstrate in Lemma 4 that the maximum amplification is $(2r+1)^2$ by showing that the existence of a distribution that can achieve a larger amplification implies in an explicit algorithm for the unstructured search problem with a larger average probability than the bound of Theorem 6. With Lemma 4, we can prove the lower bound on E_r (i.e., a general upper bound on Grover-based QAOA performance), given by Theorem 19.

Lemma 4 *For any number r of layers on Grover-based QAOA with a set S_T of ratio ρ, the amplification of the probability of measuring the elements of S_T is bounded by*

$$\eta_r(S_T, f_{X_q}) \leq (2r + 1)^2, \tag{6.10}$$

where the tight bound is achieved with $\rho \to 0$ and $q(k)$ equal to the binary function up to a scale change.

Proof: To simplify the notation in this proof, we hide the subscript with the random variable on the probability distribution and the characteristic function. In contrast, we distinguish between various probability distributions, their characteristic functions, and specific components of their summations through superscripts, without defining explicitly random variables. We also say the phase separation operator *computes* the probability distribution f if the distribution associated with $q(k)$ is f. Furthermore, to compact the notation of Eq. (6.9), we group under the notation $\Phi(\varphi, N_\varphi, k)$ both products involving characteristic functions, and we group under the notation $\mathcal{B}(N_B, k)$ the exponential factor as well as both products involving $B(\beta)$. Thus,

$$\eta_r(S_T, f) = \sum_{\boldsymbol{x}^{(L)}, \boldsymbol{x}^{(R)}} \mathcal{B}(N_B, k)\Phi(\varphi, N_\varphi, k), \tag{6.11}$$

where k is the ordered pair $(k_{\text{bra}}, k_{\text{ket}})$, N_φ is the number of characteristic functions factors, and N_B is the number of $B(\beta)$ factors.

Suppose by contradiction that for some $\epsilon > 0$, there is an choice of distribution $f^{\text{O}}(x)$ (original distribution) in which $\eta_r(S_T, f^{\text{O}}) = (2r+1)^2 + \epsilon$. We fix the optimal variational parameters. Moreover, we can set $x_o = 1$ by a shifting location without loss of generality. Note that we can express the characteristic function of $f^{\text{O}}(x)$ as

$$\varphi^{\text{O}}(\gamma) = \rho e^{i\gamma} + \varphi^{\text{rem}}(\gamma), \tag{6.12}$$

where the first term represents the portion of the summation of the characteristic function for S_T with ratio ρ and $\varphi^{\text{rem}}(\gamma)$ is the remainder portion. Let δ be a rational such that $0 < \delta \leq 2\rho$. We can rewrite $\varphi^{\text{O}}(\gamma)$ as

$$\varphi^{\text{O}}(\gamma) = 0.5\delta e^{i\gamma} + (\rho - 0.5\delta)e^{i\gamma} + \varphi^{\text{rem}}(\gamma) = 0.5\delta e^{i\gamma} + \varphi^{\text{R}}(\gamma), \tag{6.13}$$

where $\varphi^{\text{R}}(\gamma) = (\rho - 0.5\delta)e^{i\gamma} + \varphi^{\text{rem}}(\gamma)$ and $0.5\delta e^{i\gamma}$ represents the portion of the summation for a subset S_δ of S_T. Since Grover-based QAOA preserves the equality of amplitudes in degenerate states during the unitary evolution, $\eta_r(S_\delta, f^{\text{O}}) = \eta_r(S_T, f^{\text{O}})$.

Consider the following algorithm for the unstructured search problem with m marked elements over M solutions with $0.5\delta = m/M$ and r rounds. The kth diffusion operator is the sequential application of a phase separation operator that computes a target distribution $f^{\text{T}}(x)$ and the Grover mixer operator. Both with the fixed parameters of the kth layer of Grover-based QAOA. The distribution $f^{\text{T}}(x)$ has the characteristic function

$$\varphi^{\text{T}}(\gamma) = e^{i\theta\gamma}(\varphi^{\text{rea}}(\gamma) + \varphi^{\text{R}}(\gamma)). \tag{6.14}$$

The term $\varphi^{\text{rea}}(\gamma)$ represents an arbitrary reassignment of the costs of the marked elements of the original distribution replacing $0.5\delta e^{i\gamma}$ and the factor $e^{i\theta\gamma}$ is a location shift of the distribution of size $\theta > 0$. The quantity m can be chosen as the minimum number of marked elements required for computing the distribution $f^{\text{T}}(x)$.

On the other hand, for marked elements, the oracle reverses the action of the defined phase separation and then applies a phase shift of $e^{-i\gamma}$, i.e., a mapping on the cost 1. In practice, the oracle interrupts the computation of target distribution $f^{\text{T}}(x)$ of the phase separation on specific marked values in such a way that the combined action of diffusion and oracle operators encodes the value 1 (note that location shift of $e^{i\theta\gamma}$ was introduced so that just marked elements be mapped on the cost 1). Consequently, the algorithm's performance depends on the positions of the marked elements, and thereby the average probability is unknown. For using the optimality of Theorem 6, we bound the minimum probability value in terms of the known performance of original distribution $f^{\text{O}}(x)$ by choosing values of δ and θ sufficiently small.

Let $f^{\text{sp}}(x)$ (search problem) be a distribution computed by the combined action of the phase separation and the oracle for an arbitrary instance of the search problem algorithm. The characteristic function of $f^{\text{sp}}(x)$ can be expressed expressed without loss of generality by

$$\varphi^{\text{sp}}(\gamma) = 0.5\delta e^{i\gamma} + e^{i\theta\gamma}\varphi^{\text{R}}(\gamma) + \varphi^{1}(\gamma) - \varphi^{2}(\gamma), \tag{6.15}$$

where $0.5\delta e^{i\gamma}$ represents the oracle finding the marked elements; $e^{i\theta\gamma}\varphi^{\text{R}}(\gamma)$ is the portion of $f^{\text{T}}(x)$ computed up to the phase shifting $e^{i\theta\gamma}$ on the original characteristic function $\varphi^{\text{O}}(x)$; $\varphi^{1}(\gamma)$ represents the non-computed part of $f^{\text{T}}(x)$ on the original distribution $f^{\text{O}}(x)$, that is computed with on distribution $f^{\text{sp}}(x)$; and $\varphi^{2}(\gamma)$, in contrast, represents the non-computed part of $f^{\text{T}}(x)$ on the distribution $f^{\text{sp}}(x)$, that is computed on $f^{\text{O}}(x)$. We denote

$$\varphi^{\text{eq}}(\gamma) = 0.5\delta e^{i\gamma} + e^{i\theta\gamma}\varphi^{\text{R}}(\gamma),\ \varphi^{\text{dif}}(\gamma) = \varphi^{1}(\gamma) - \varphi^{2}(\gamma). \tag{6.16}$$

The first definition, $\varphi^{\text{eq}}(\gamma)$ (equal), represents the original distribution $f^{\text{O}}(x)$ up to the location shifting on $f^{\text{R}}(x)$. The second definition, $\varphi^{\text{dif}}(\gamma)$ (different), represents the divergence between positions of $f^{\text{O}}(x)$ and $f^{\text{sp}}(x)$. If the marked elements on both distributions are at the same positions, $\varphi^{\text{dif}}(\gamma) = 0$, and since at worst case it diverges on all their $2m$ marked elements, $|\varphi^{\text{dif}}(\gamma)| \leq \delta$. By Eq. (6.13) and (6.16), $\varphi^{\text{eq}}(\gamma)$ can be written as $\varphi^{\text{eq}}(\gamma) = \varphi^{\text{O}}(\gamma) + (e^{i\theta\gamma} - 1)\varphi^{\text{R}}(\gamma)$ and then we can write $\varphi^{\text{sp}}(\gamma)$ from Eq. (6.15) in terms of $\varphi^{\text{O}}(\gamma)$ by

$$\varphi^{\text{sp}}(\gamma) = \varphi^{\text{O}}(\gamma) + (e^{i\theta\gamma} - 1)\varphi^{\text{R}}(\gamma) + \varphi^{\text{dif}}(\gamma). \tag{6.17}$$

Using $\gamma^{(j)}$ as a generic notation for the argument of the jth characteristic function on the product (for any arbitrary order) and setting $\vartheta = \theta\gamma_{max}$ with γ_{max} being the maximum absolute value of an argument, we bound $\Phi(\varphi^{\mathrm{sp}}, N_\varphi, k)$ by

$$\begin{aligned}
&\Phi(\varphi^{\mathrm{sp}}, N_\varphi, k) = \prod_{j=1}^{N_\varphi} \varphi^{\mathrm{sp}}(\gamma^{(j)}) \\
&= \prod_{j=1}^{N_\varphi} \varphi^{\mathrm{O}}(\gamma^{(j)}) + ((e^{i\theta\gamma^{(j)}} - 1)\varphi^{\mathrm{R}}(\gamma^{(j)}) + \varphi^{\mathrm{dif}}(\gamma^{(j)})) \\
&= \sum_{\boldsymbol{x}} \prod_{j=1}^{N_\varphi} (\varphi^{\mathrm{O}}(\gamma^{(j)}))^{1-x_j} ((e^{i\theta\gamma^{(j)}} - 1)\varphi^{\mathrm{R}}(\gamma^{(j)}) + \varphi^{\mathrm{dif}}(\gamma^{(j)}))^{x_j} \\
&= \prod_{j=1}^{N_\varphi} \varphi^{\mathrm{O}}(\gamma^{(j)}) + \sum_{\boldsymbol{x} \smallsetminus \mathbf{0}} \prod_{j=1}^{N_\varphi} (\varphi^{\mathrm{O}}(\gamma^{(j)}))^{1-x_j} ((e^{i\theta\gamma^{(j)}} - 1)\varphi^{\mathrm{R}}(\gamma^{(j)}) + \varphi^{\mathrm{dif}}(\gamma^{(j)}))^{x_j} \\
&\geq \Phi(\varphi^{\mathrm{O}}, N_\varphi, k) - 2^{N_\varphi}(\delta + \vartheta),
\end{aligned} \tag{6.18}$$

where $\boldsymbol{x} = (x_1, \ldots, x_{N_\varphi})$ is a N_φ-bit string and $\mathbf{0}$ is a vector of N_φ zeros. The inequality follows from the individual bounds $|\varphi^{\mathrm{O}}(\gamma^{(j)})| \leq 1$, $|\varphi^{\mathrm{R}}(\gamma^{(j)})| \leq 1$, $|\varphi^{\mathrm{dif}}(\gamma^{(j)})| \leq \delta$, and

$$|e^{i\theta\gamma^{(j)}} - 1| = \sqrt{2}\sqrt{1 - \cos(\theta\gamma^{(j)})} \leq \theta\gamma^{(j)} \leq \vartheta, \tag{6.19}$$

and the fact that there is at least one j in each term of the summation in which $x^{(j)} = 1$. The first inequality of Eq. (6.19) follows from $\cos(x) \geq 1 - x^2/2$. The maximum value of both N_φ and N_B are $2r$ and $|\mathcal{B}(N_B, k)|$, which is equal for both $f^{\mathrm{O}}(x)$ and $f^{\mathrm{sp}}(x)$ distributions since we fix the parameters, is bounded by 2^{N_B}. Combining those results with Eq. (6.11) gives

$$\eta_r(S_\delta, f^{\mathrm{sp}}) \geq \eta_r(S_\delta, f^{\mathrm{O}}) - 64^r(\delta + \vartheta) = (2r + 1)^2 + \epsilon - 64^r(\delta + \vartheta). \tag{6.20}$$

For any r, there is a choices of δ and ϑ in which $\epsilon > 64^r(\delta + \vartheta)$. Combining it with the fact that the maximum amplification on Grover's algorithm is $(2r + 1)^2$, the optimality of Theorem 6 is contradicted and establishes the lemma. □

Theorem 19 *For any number r of layers in Grover-based QAOA, the expectation value is bounded by*

$$E_r \geq G_{X_c}(\tau_1)(2r + 1)^2 + \tau_2(1 - F_{X_c}(\tau_1)(2r + 1)^2), \tag{6.21}$$

where τ_1 is the maximum element of the support of X_c in which $F_{X_c}(t) \leq 1/(2r+1)^2$ and τ_2 is the minimum element in which $F_{X_c}(t) > 1/(2r + 1)^2$. In particular, if $F_{X_c}(\tau_1) = 1/(2r + 1)^2$, then $E_r \geq \mathrm{E}[X_c | X_c \leq \tau_1]$.

Proof: To build our upper bound on expectation value, we assume the largest amplification of $(2r+1)^2$, bounded by Lemma 4, for the smallest solutions until τ_1. The remainder probability is assigned to τ_2. The expectation value E_r is bounded with a weighted sum of the expectation values of X_c given $X_c \leq \tau_1$ and X_c given $X_c = \tau_2$ by its amplified probabilities. Thus,

$$\begin{aligned} E_1 &\geq \mathrm{E}[X_c|X_c \leq \tau_1]F_{X_c}(\tau_1)(2r+1)^2 + \mathrm{E}[X_c|X_c = \tau_2](1 - F_{X_c}(\tau_1)(2r+1)^2) \\ &= G_{X_c}(\tau_1)(2r+1)^2 + \tau_2(1 - F_{X_c}(\tau_1)(2r+1)^2). \end{aligned} \tag{6.22}$$

If $F_{X_c}(\tau_1) = 1/(2r+1)^2$, the second term vanishes and $E_r \geq \mathrm{E}[X_c|X_c \leq \tau_1]$. □

The equality of Eq. (6.21) is referred to as the *maximum amplification bound.* The bound is not tight since we can reach probability 1 on the search problem only if ρ is at least the larger ratio of $\rho_{Th}(r)$, a consequence of the fact that the amplification decreases as we move away from the low-convergence regime. For instance, we need to a ratio of $\rho = 0.25$ to achieve probability 1 on $r = 1$, instead of $\rho = 1/(2r+1)^2 = 1/9$. Note that the MAOA operates close to the regime of the maximum amplification, although it does not use the expectation value as a metric. Despite this, the maximum amplification has the same asymptotic behavior as GM-Th-QAOA in all aspects considered—as a result, the same asymptotic behavior emphasized in the numerical experiments of Chapter 5 could be reached by computing the maximum amplification bound. Firstly, if X is continuous, $F_{X_c}(\tau_1) = 1/(2r+1)^2$ for any r and the bound $E_r \geq \mathrm{E}[X_c|X_c \leq \tau_1]$ combined with $F_{X_c}(\tau_1) = \Theta(1/r^2)$ gives Corollary 11, a generalization of Theorem 15 which follows using analogous arguments.

Corollary 11 *For Grover-based QAOA, if X_c is a continuous distribution and $f_{X_c}(R_{X_c}^{min}) = a$, where $0 < a < \infty$, then the quantile achieved by the expectation value is asymptotically bounded by*

$$F_{X_c}(E_r) = \Omega\left(\frac{1}{r^2}\right). \tag{6.23}$$

Corollary 11 implies that any Grover-based QAOA cannot be asymptotic better than the quadratic Grover-like speed-up, confirming the motivation for the discussion of this chapter and establishing the most important conclusion of this dissertation. Moreover, all the constructions of Subsec. 4.2.4 are applicable to the maximum amplification bound. Now, combining Corollary 7 and Theorem 19 and assuming X continuous gives a comparison of GM-Th-QAOA with maximum amplification bound on the large limit of r. Using the definition of L of Eq. (4.85) on Eq. (4.75) gives $F_{X_c}(E_r(t)) \leq \frac{L\pi^2}{16r^2}$ and on Eq. (6.21), $F_{X_c}(E_r(t)) \geq \frac{L}{4r^2}$. Therefore,

$$\frac{L}{4r^2} \leq F_{X_c}(E_r(t)) \leq \frac{L\pi^2}{16r^2}, \tag{6.24}$$

and thus GM-Th-QAOA is, in the worst case, $\pi^2/4$ times worse than the maximum amplification bound in terms of the cdf.

With the maximum amplification bound, we can also bound $C(r)$—and consequently on the number of rounds to achieve a fixed approximation ratio—obtaining the analogous of Theorem 16 and Corollary 10 for Grover-based QAOA, synthesized in Theorem 20. In a similar way to GM-Th-QAOA, the proof consists of concluding that the binary function achieves the maximum C_r, but now by slightly modifying the argument of Lemma 3.

Theorem 20 *For any number r of layers in Grover-based QAOA, $C(r) \leq 2\sqrt{r(r+1)}$ and, provided that $R_X^{min} \neq 0$ and $|R_X^{min}| < \infty$,*

$$r \geq \frac{\mu - \lambda R_{X_c}^{min}}{2\sigma\sqrt{1 + 1/r}}. \tag{6.25}$$

Proof: Consider τ_1 and τ_2 from Theorem 19. Let X_1 and X_2 be random variables where the probability distribution $f_{X_1}(x)$ is given by $f_{X_1}(x) = f_{X_c}(x)$ for all $x \neq \tau_2$ on the support R_{X_c}, $f_{X_1}(\tau_2) = 1/(2r+1)^2 - F_{X_c}(\tau_1)$, and the remainder probability to reach the summation of probabilities equal to 1 can be arbitrarily assigned on values above τ_2; and the probability distribution $f_{X_2}(x)$ is $f_{X_2}(x) = f_{X_c}(x)$ for all $x \neq \tau_2$ on R_{X_c}, $f_{X_2}(\tau_2) = f_{X_c}(\tau_2) - f_{X_1}(\tau_2)$ with the remainder probability again arbitrarily assigned but for values below τ_2.

We can think of X_1 and X_2 as a partition of the probability of the value τ_2 of the distribution X_c preserving the remainder values of R_{X_c}. The first takes the exact probability necessary to complete $1/(2r+1)^2$, i.e., $1/(2r+1)^2 - F_{X_c}(\tau_1)$, while the last takes the remainder $f_{X_c}(\tau_2) - f_{X_1}(\tau_2)$. Consequently, X_c is connected with random variables X_1 and X_2 by the property that for any function $f(x)$, from LOTUS,

$$\begin{aligned} \mathrm{E}[f(X_c)] &= \sum_{x \in R_{X_c}} f(x) f_{X_c}(x) \\ &= \left(\sum_{x \in R_{X_1} : x \leq \tau_2} f(x) f_{X_1}(x) \right) + \left(\sum_{x \in R_{X_2} : x \geq \tau_2} f(x) f_{X_2}(x) \right). \end{aligned} \tag{6.26}$$

.

Furthermore, the random variable X_1 allows us to write the bound of Theorem 19 as

$$E_r \geq \mathrm{E}[X_1 | X_1 \leq \tau_2] = \frac{G_{X_1}(\tau_2)}{F_{X_1}(\tau_2)}, \tag{6.27}$$

since we choose the probability for the value τ_2 such that $F_{X_1}(\tau_2) = 1/(2r+1)^2$. We can assume, without loss of generality, that $\mu = 0$ since a location shift does not affect C_r. Following the analogous step of Lemma 3, we split σ^2 summation into

$$\sigma^2_{\leq\tau_2} = \sum_{x \in R_{X_1} : x \leq \tau_2} x^2 f_{X_1}(x), \; \sigma^2_{\geq\tau_2} = \sum_{x \in R_{X_2} : x \geq \tau_2} x^2 f_{X_2}(x), \tag{6.28}$$

where $\sigma = \sqrt{\sigma^2_{\leq\tau_2} + \sigma^2_{\geq\tau_2}}$. Let $X_{\leq\tau_2}$ be the random variable X_1 given $X_1 \leq \tau_2$ and $X_{\geq\tau_2}$ the random variable X_2 given $X_2 \geq \tau_2$. By Eq. (6.26),

$$\begin{aligned} \mathrm{E}[X_{\leq\tau_2}] &= \frac{G_{X_1}(\tau_2)}{F_{X_1}(\tau_2)}, \; \mathrm{E}[X^2_{\leq\tau_2}] = \frac{\sigma^2_{\leq\tau_2}}{F_{X_1}(\tau_2)}, \\ \mathrm{E}[X_{\geq\tau_2}] &= -\frac{G_{X_1}(\tau_2)}{1 - F_{X_1}(\tau_2)}, \; \mathrm{E}[X^2_{\geq\tau_2}] = \frac{\sigma^2_{\geq\tau_2}}{1 - F_{X_1}(\tau_2)}, \end{aligned} \tag{6.29}$$

which results in

$$\frac{|G_{X_1}(\tau_2)|}{\sigma} \leq \sqrt{F_{X_1}(\tau_2)(1 - F_{X_1}(\tau_2))}. \tag{6.30}$$

Combining it with Eq. (6.27) gives

$$\begin{aligned} C_r \leq & \frac{\sqrt{F_{X_1}(\tau_2)(1 - F_{X_1}(\tau_2))}}{F_{X_1}(\tau_2)} = \sqrt{\frac{1}{F_{X_1}(\tau_2)} - 1} = \sqrt{(2r+1)^2 - 1} \\ & = 2\sqrt{r(r+1)}, \end{aligned} \tag{6.31}$$

as desired. The bound of Eq. (6.25) follows from

$$r \geq \frac{\mu - \lambda R^{min}_{X_c}}{(C_r/r)\sigma}, \tag{6.32}$$

and $C_r/r \leq 2\sqrt{1 + 1/r}$. □

The bound on the quantity $C(r)/r$ is decreasing in r, with a maximum of $2\sqrt{2}$ in $r = 1$ and a minimum of 2 in $r \to \infty$. Combining Theorems 16 and 20, $\kappa r \leq C^{GM}(r) \leq 2r$ on large r. Note that Theorem 20 improve Benchasattabuse et al. [122] bound of Eq. (3.34) by a constant factor of $\sqrt{2}\pi$ on $r = 1$ and 2π on $r \to \infty$. Beyond the more general context of Grover-based QAOA, our lower bound has the advantage of allowing any cost function instead of only cost functions with non-positive integer costs. We get also the analogous of the Eq. (4.105) and (4.106) for the number of rounds to reached probability 1 of measuring a optimal solution with

$$f_{X_c}(R^{min}_{X_c}) \geq \frac{1}{(2r+1)^2} \;\Rightarrow\; r \geq \frac{1}{2}\left(\frac{1}{\sqrt{f_{X_c}(R^{min}_{X_c})}} - 1\right), \tag{6.33}$$

and

$$r \geq \frac{1}{2\sqrt{f_{X_c}(R_{X_c}^{min})}} = \Omega\left(\frac{1}{\sqrt{f_{X_c}(R_{X_c}^{min})}}\right) \tag{6.34}$$

as $f_{X_c}(R_{X_c}^{min}) \to 0$, respectively.

To finish, a direct comparison with Grover Adaptive Search concerning exact optimization follows directly from the bound on amplification of Lemma 4. Since the probability is bounded by $f_{X_c}(R_{X_c}^{min})(2r+1)^2$, finding an optimal solution for an optimization problem with Grover-based QAOA with probability as least 1/2 needs $\Omega(1/\sqrt{f_{X_c}(R_{X_c}^{min})})$ rounds, the analog complexity of GAS.

6.3 Bounds on Max-Cut

One application of our bounds is the Max-Cut problem. Firstly, we can improve by a constant factor the Max-Cut lower bound of Benchasattabuse et al. [122] on bipartite graphs, given by Eq. (3.35), with

$$r \geq \frac{2\lambda - 1}{2\sqrt{1+1/r}}\sqrt{|\mathcal{E}|}. \tag{6.35}$$

The analogous bound with GM-Th-QAOA, by Corollary 10 on large limit of r, is

$$r \geq \frac{2\lambda - 1}{\kappa}\sqrt{|\mathcal{E}|}. \tag{6.36}$$

In general, if the statistical quantity $(\mu - R_{X_c}^{min})/\sigma$ grows with the size of the instance to a given COP, we cannot achieve a fixed approximation ratio with a constant number of layers, which, as emphasized on Sec. 3.11, is a severe limitation for the NISQ context. However, at least in Max-Cut, the situation seems to be even worse, as we can see applying the bound of Eq. (6.33) and (6.34).

In that case, we consider additionally that the bipartite graph is connected. By the well-known fact that these graphs have a unique bipartition and that in the binary codification of Max-Cut on QAOA the solutions are duplicated, the number of cuts of maximum size is 2 and thus $f_{X_c}(R_{X_c}^{min}) = 1/2^{|\mathcal{V}|-1}$. Therefore, Eq. (6.33) and (6.34) give

$$r \geq \frac{1}{2}\left(2^{\frac{|\mathcal{V}|-1}{2}} - 1\right) \Rightarrow r \geq 2^{\frac{|\mathcal{V}|-3}{2}} = \Omega(\sqrt{2^{|\mathcal{V}|}}) \text{ as } |\mathcal{V}| \to \infty. \tag{6.37}$$

Doing the same for GM-Th-QAOA with Eq. (4.105) and (4.106), we get

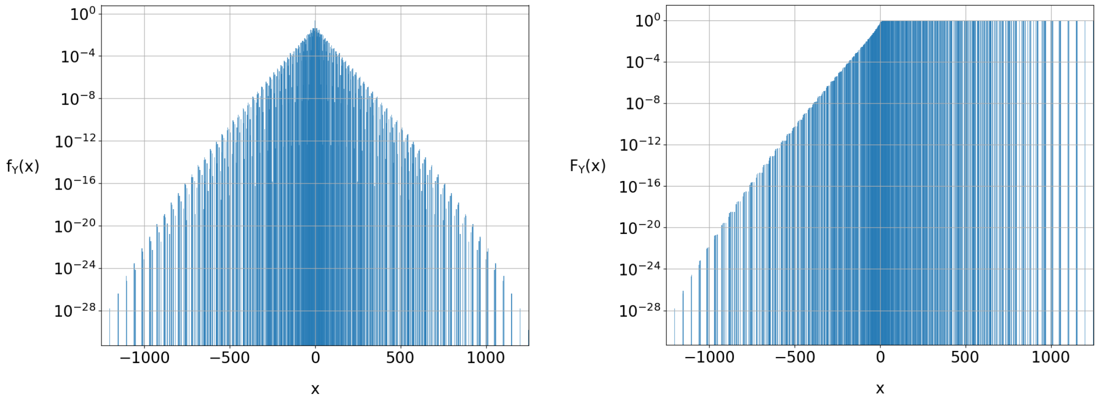
$$r \geq \frac{1}{4}\left(\frac{\pi}{\arcsin\left(1/\sqrt{2^{|\mathcal{V}|-1}}\right)} - 2\right) \Rightarrow r \geq \pi\, 2^{\frac{|\mathcal{V}|-5}{2}} = \Omega(\sqrt{2^{|\mathcal{V}|}}) \text{ as } |\mathcal{V}| \to \infty. \tag{6.38}$$

Both scale exponentially with the number of vertices. As connected graphs have at least $|\mathcal{V}| - 1$ edges, it scales exponentially also with the number of edges. Of course, the bound is not applicable on approximate solutions with $\lambda < 1$. Nevertheless, at least for the class of the complete bipartite graphs, we can argue that the growth is exponential by analyzing its probability distribution.

Let us consider the complete bipartite graph $K_{n,n}$ with bipartition on the sets V_1 and V_2. Suppose a solution of Max-Cut as a partition on the sets S_1 and S_2 in which among the vertices of S_1, the number of vertices that belong to V_1 and V_2 are respectively j and k. Note that an edge (u,v) of $K_{n,n}$ does not belong to the cut if and only if both vertices u and v are at the same cut partition (S_1 or S_2). Therefore, the size of the cut can be computed by discounting to the total number of edges n^2 of the whole graph, $jk + (n-j)(n-k)$, i.e., the number of edges induced by the union of both complete bipartite graph $K_{j,k}$ and $K_{n-j,n-k}$ within the sets S_1 and S_2, respectively. Thus, with our definition of considering minimization problems and using the random variable Y by subtraction of the mean $-n^2/2$, the cost of the solution is $\frac{1}{2}(n-2j)(n-2k)$. The number of solutions to a given j and k is count by choosing j between n vertices of V_1 and k between n vertices of V_2, resulting in $\binom{n}{j}\binom{n}{k}$. That way, we can characterize the solution space by

$$\binom{n}{j}\binom{n}{k} \text{ solutions of cost } \frac{1}{2}(n-2j)(n-2k) \text{ for all } 0 \leq j,k \leq n. \tag{6.39}$$

Although we do not have the explicit distribution, once, in general, there are different combined choices of j and k with the same cost, this is sufficient to see that it presents an exponential decay toward the optimal solution. To get that, note that if we fix j and go through all the values of k, the induced function is a multiple of a symmetric binomial distribution with a change on the scale. By CLT, we know that binomial distribution gets closer to a normal distribution with large n, which explicitly has an exponential decay in the tails. By symmetric, the same happens by fixing k and varying j. The combined behavior implies a trend of a linear increase of the costs to lower values of k and j, accompanied by an exponential decay of their probabilities. Fig. 6.1 illustrates the decay of the distribution for the graph $K_{50,50}$ by showing the graphics of $f_Y(x)$ and $F_Y(x)$. The exponential decay, with the arguments of the Subsec. 4.2.4 on the asymptotic limit, infers in a logarithmic increase of C_r with the number of layers, and therefore, as $(\mu - R_{X_c}^{min})/\sigma$ grows with the square root on the number of edges for bipartite graphs, the number of layers to a achieves a fixed λ must increase exponentially with the number of vertices/edges.

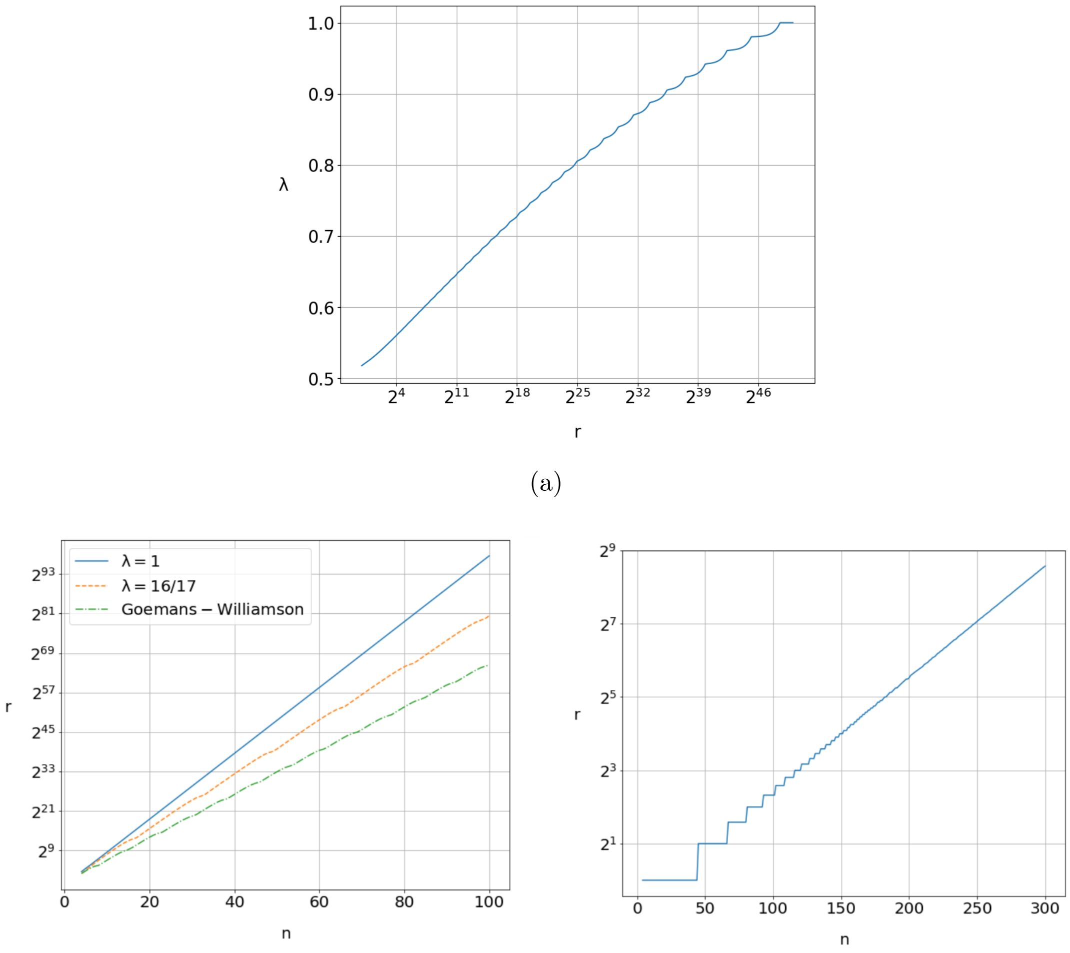


(a) (b)

Figure 6.1: (a) Probability mass function and (b) cumulative distribution function concerning the random variable Y for the Max-Cut problem on the graph $K_{50,50}$. We compute efficiently the explicit distribution from our characterization of the solution space given by Eq. (6.39). Both graphs are on a log-linear scale. As the possible values for the cost are given by $\frac{1}{2}(n-2j)(n-2k)$ for all $0 \leq j, k \leq n$, the region near to 0 is denser than the regions near to $R_{X_c}^{min}$ and R_X^{max}. Although the decay is not uniform, the general trend is clearly exponential.

To illustrate our arguments, we simulate the maximum amplification bound on the complete bipartite graphs of different sizes. Fig. 6.2(a) shows the logarithm growth of the approximation ratio on r to the graph $K_{50,50}$. Fig. 6.2(b) and Fig. 6.2(c) display the exponential dependence on the number of rounds to achieve different values of approximation ratio when we scale n. In the first between the last two figures mentioned, we choose approximation ratios with practical interest, while in the last one, we consider an extremely low approximation ratio to emphasize the stiffness of the limitation.

In a more general context, a given type of instance must suffer from the same limitation if there is simultaneously a distribution with exponential decay and the quantity $(\mu - R_{X_c}^{min})/\sigma$ grows above the logarithmic rate with the size of the problem. Thus, it is likely that there are other classes of graphs on Max-Cut and other types of instances beyond the Max-Cut problem that fit into these conditions. For example, we can mention the aforementioned normally distributed instances of the Capacitated Vehicle Routing and Portfolio Optimization problems [23, 24], which meet the first criterion.



(a)

(b)

(c)

Figure 6.2: (a) Linear-log base 2 plot of approximation ratio versus the number of layers considering the application of the maximum amplification bound for the graph $K_{50,50}$ on Max-Cut. As expected, the growth rate is logarithmic. The resolution used on the number of layers (due to the extremely high number of layers to get $\lambda = 1$) is $\lceil 2^{x/100} \rceil$ for $x = 0, 1, \ldots, 5000$. Since the probability distribution is discrete, the values τ_1 and τ_2 on Eq. (6.21) were obtained via testing the point of the support via brute force. (b) The minimum number of layers required to maximum amplification bound achieves three different values of approximation ratio on the Max-Cut with the graph $K_{n,n}$ for $n = 4, 5, \ldots, 100$. The considered values of λ are $\lambda = 1$, in which we calculate analytically that $r = \lceil 2^{0.5(|\mathcal{V}|-3)} \rceil$; $\lambda = 16/17 \approx 0.9412$, the λ value in which Max-Cut becomes NP-Hard; and the approximation ratio guaranteed by the classical Goemans-Williamson algorithm, given by $\lambda \approx 0.8786$. The scale is log-linear with base 2. The value of r in which the approximation ratio is achieved was efficiently found with a binary search. As predicted, the number of layers scales exponentially in all of them. (c) The same as (b), but with $n = 4, 5, \ldots, 300$ and $\lambda = 0.52$. As the expectation value of a uniform superposition gives $\lambda = 0.5$, this approximation ratio is extremely low. However, even for such low performance, given a sufficient number of vertices, we observe the exponential dependence on the number of layers r.

Chapter 7

Conclusion

In the present work, we first develop independently a similar statistical approach to the Headley and Wilhelm [30] paper for GM-QAOA, obtaining the equivalent expression for expectation value of complexity $\mathcal{O}(4^r)$. Although the method is insightful and allows obtaining angles up to almost something close to ten rounds, the expression is a dead end for obtaining theoretical bound on the performance through direct analytical treatment that contributed to the discussion of Sec. 3.11, especially for the issue of whether the Grover mixer variants of QAOA are limited to quadratic Grover-style speed-up. To bypass that issue, we extend the statistical approach to GM-Th-QAOA, which is simpler due to its binary phase separation operator, taking advantage of the optimality of Grover's algorithm on the unstructured search to obtaining an expression for the expectation value with complexity independent of the number of layers. With the expression, we first solve the threshold curve problem and then get bounds of different natures, including on the statistical quantities of quantile and the standard score, and on the minimum number of layers required to get a fixed approximation ratio. The bound on the quantile is of particular interest since it reflects explicitly a quadratic Grover-style speed-up. Subsequently, we generalize the GM-Th-QAOA bounds to the general Grover-based QAOA framework, by using an indirect argument with the optimality of the unstructured search problem, achieving the same asymptotic performance and obtaining the main contribution of this work, that is, the formal establishment that the Grover mixer has the performance bounded by the quadratic bound of the unstructured search problem, i.e., a quadratic Grover-style speedup over classical brute force. That limiting can be severe for combinatorial optimization, as evidenced by the application of Max-Cut on the complete bipartite graph, which requires an exponential number of layers to maintain constant performance.

In this way, to get significant results with QAOA, especially in the NISQ era, it is essential for the algorithm to explore the structure of the optimization problems. Indeed, recall the numerical evidence of Golden et al. [11] that suggests the possibility of exponential gain of QAOA with structure-dependent mixer over Grover mixer variants. Thus, research should be directed toward understanding the mechanisms by which different types of mixers can benefit from the structure of particular problems, a path opened by Headley [33] with the statistical approach on the transverse field mixer and the line mixer.

These contributions are significant within a broader context. Despite the QAOA being the most prominent quantum algorithm for combinatorial optimization, much of our knowledge of its general performance is heuristic. Thus, this work signifies a pivotal step towards a solid understanding of the performance and limitations of this class of algorithms, consequently providing insights into the potential of quantum computing for tackling combinatorial optimization problems.

Yet in the Grover mixer context, there are still open questions and paths to exploit.

- Decide whether GM-Th-QAOA is the best Grover-based QAOA for all possible instances, or at least whether GM-Th-QAOA outperforms GM-QAOA always, confirming the numerical evidence (of the previous literature and of this work). Intuitively, it is reasonable to think that the most efficient agnostic-structure method possible is to compile the cost function on a binary function and perform Grover's algorithm. The results and insights of the present work indicate that this can be the case. However, formal proof is still needed;

- Decide whether GM-QAOA even reflected the quadratic Grover-like speed-up in the sense of Theorems 15 and 17 or with another metric. However, insights would be needed to answer that question analytically since direct analytical treatment to bound the equations of Theorem 18 is infeasible. In an empirical sense, as mentioned in the introduction of Chapter 1, the recent work of Zhang et al. [41] provides numerical evidence of the quadratic speed-up. In particular, that work studied a version of the Satisfiability problem in which every clause has exact 3 literals, called 3-SAT. That decision problem was translated into QAOA language as a Max-3-SAT, and a metric of performance considered was the number of rounds to reach at least a probability 0.5 of measuring a quantum state corresponding to the solution of the original 3-SAT decision problem. Denoting by p the ratio of assignments that are solutions, the numerical experiments on several random instances considering the range of $n = 10, \dots, 26$ boolean variables indicate that the aforementioned metric scales as $1/\sqrt{p}$, which is a quadratic speed-up over CRS;

- Decide whether the limit L of Eq. (4.85) is finite non-zero for any continuous probability probability, relaxing the hypothesis of $f_X(R_X^{min}) = a$, where $0 < a < \infty$ on Theorem 15 if the answer is yes;

- The application of the maximum amplification bound to more graph classes on Max-Cut and other combinatorial optimization problems would be welcome. This way, we could know how common is the need for exponential growth on the layers. Fortunately, explicit knowledge of the distribution is not necessarily required to find the asymptotic behavior and establish how the performance scales, as was the case of complete bipartite graphs on Max-Cut;

- The numerical experiments for families of probability distribution in Chapter 5 were considerably comprehensive. However, more cases could be investigated. For example, the instances of Capacitated Vehicle Routing observed by Bennett and Wang [23] are not normally perfectly distributed, presenting asymmetry. That way, it would be convenient to consider the skew normal [129–131] distributions, a class of distributions that generalize normal distribution by adding shape parameters related to the skewness. It would also be interesting to consider relaxing the hypothesis of a finite second moment to consider distributions such as Pareto(α, x_m) with $\alpha \leq 2$. Another family of distributions worth mentioning is distributions with 3 points. Since we proved that the maximum standard score is achieved by distributions with 2 points, it would be intriguing to see what happens by adding a third point. Finally, we can cite traditional distributions of the literature, such as beta, Poisson, geometric, and negative binomial distributions [31];

- Develop statistical methods to determine QAOA angles for combinatorial optimization problems through numerical experimentation with probability distributions—such as the numerical experiments of Chapter 5. In an analytical sense, Headley and Wilhelm [30] found the distribution of the Number Partition Problem. On the other hand, Marsh and Wang [23] observed normally distributed instances of Capacitated Vehicle Routing and Portfolio Optimization problems. Naturally, one can think of a method for parameter estimation of GM-QAOA based on the assumption of the solution space normally distributed for problems, such as Capacitated Vehicle Routing, and the obtaining of estimators with a random sample of efficient size with a fixed confidence interval for the standard deviation—on the distribution Normal(u, s^2), the angles $\boldsymbol{\beta}$ are independent of u and s, while the angles $\boldsymbol{\gamma}$ depends only on s (inversely proportional to the standard deviation, according to Corollary 3). More refined methods would model the problem with a skew normal distribution, considering an estimator for the skewness.

  In general, for problems with probability distribution unknown and weakly known, one can develop a heuristic method based on efficient estimators for sample moments. For instance, in addition to the standard deviation, we can truncate the expansion of Eq. (4.48) until moments of a given order, such as kurtosis or higher, and with the estimators of these moments, plugging the approximation of Eq. (4.48) on Eq. (4.47). The obtained angles probably would not be accurate enough to get desirable results by the direct sampling of a quantum circuit, but at least maybe there could be a better parameter initialization for the outer loop of QAOA than the random initialization.

However, we must emphasize, as concluded from the results of this work, that GM-QAOA may not be a suitable algorithm for combinatorial optimization, having lower performance than GM-Th-QAOA itself and exhibiting, at most, quadratic speed-up over the classical brute force. Despite this, as indicated by Headley [33] with the generalization of the statistical approach to the transverse field and line mixer, the estimator of statistical quantities of probability distributions is a promising path in the studies of angles finding of QAOA. That way, GM-QAOA can be a starting point for obtaining proof of concepts for statistical methods, having in your favor the classical simulation discussed in Subsec. 3.8.3, which allows simulates larger instances than the transverse field mixer on the current state-of-art.

References

[1] NIELSEN, M. A., CHUANG, I. L. *Quantum computation and quantum information.* 10 ed. Cambridge, Cambridge University Press, 2010.

[2] PRESKILL, J. "Quantum computing in the NISQ era and beyond", *Quantum*, v. 2, pp. 79, 2018.

[3] CHENG, B., DENG, X.-H., GU, X., et al. "Noisy intermediate-scale quantum computers", *Frontiers of Physics*, v. 18, n. 2, pp. 21308, 2023.

[4] CEREZO, M., ARRASMITH, A., BABBUSH, R., et al. "Variational quantum algorithms", *Nature Reviews Physics*, v. 3, n. 9, pp. 625–644, 2021.

[5] FARHI, E., GOLDSTONE, J., GUTMANN, S. "A quantum approximate optimization algorithm", *arXiv preprint arXiv:1411.4028*, 2014.

[6] HADFIELD, S., WANG, Z., O'GORMAN, B., et al. "From the quantum approximate optimization algorithm to a quantum alternating operator ansatz", *Algorithms*, v. 12, n. 2, pp. 34, 2019.

[7] FARHI, E., GOLDSTONE, J., GUTMANN, S., et al. "Quantum computation by adiabatic evolution", *arXiv preprint quant-ph/0001106*, 2000.

[8] FARHI, E., GOLDSTONE, J., GUTMANN, S., et al. "A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem", *Science*, v. 292, n. 5516, pp. 472–475, 2001.

[9] BERNHARD, K., VYGEN, J. *Combinatorial optimization: Theory and algorithms.* 5 ed. Berlim, Springer-Verlag, 2012.

[10] BLEKOS, K., BRAND, D., CESCHINI, A., et al. "A review on quantum approximate optimization algorithm and its variants", *Physics Reports*, v. 1068, pp. 1–66, 2024.

[11] GOLDEN, J., BÄRTSCHI, A., O'MALLEY, D., et al. "Numerical evidence for exponential speed-up of QAOA over unstructured search for approximate constrained optimization". In: *2023 IEEE International Conference on*

Quantum Computing and Engineering (QCE), v. 1, pp. 496–505. IEEE, 2023.

[12] MARSH, S., WANG, J. B. "Combinatorial optimization via highly efficient quantum walks", *Physical Review Research*, v. 2, n. 2, pp. 023302, 2020.

[13] BÄRTSCHI, A., EIDENBENZ, S. "Grover mixers for QAOA: Shifting complexity from mixer design to state preparation". In: *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 72–82. IEEE, 2020.

[14] MARSH, S., WANG, J. B. "A quantum walk-assisted approximate algorithm for bounded NP optimisation problems", *Quantum Information Processing*, v. 18, pp. 1–18, 2019.

[15] FUCHS, F. G., LYE, K. O., MØLL NILSEN, H., et al. "Constraint preserving mixers for the quantum approximate optimization algorithm", *Algorithms*, v. 15, n. 6, pp. 202, 2022.

[16] AKSHAY, V., PHILATHONG, H., MORALES, M. E., et al. "Reachability deficits in quantum approximate optimization", *Physical Review Letters*, v. 124, n. 9, pp. 090504, 2020.

[17] GOLDEN, J., BÄRTSCHI, A., O'MALLEY, D., et al. "The quantum alternating operator ansatz for satisfiability problems". In: *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)*, v. 1, pp. 307–312. IEEE, 2023.

[18] GROVER, L. K. "A fast quantum mechanical algorithm for database search". In: *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pp. 212–219, 1996.

[19] GROVER, L. K. "Quantum mechanics helps in searching for a needle in a haystack", *Physical Review Letters*, v. 79, n. 2, pp. 325, 1997.

[20] FARHI, E., GUTMANN, S. "Quantum computation and decision trees", *Physical Review A*, v. 58, n. 2, pp. 915, 1998.

[21] KEMPE, J. "Quantum random walks: an introductory overview", *Contemporary Physics*, v. 44, n. 4, pp. 307–327, 2003.

[22] PORTUGAL, R. *Quantum walks and search algorithms*. 2 ed. Cham, Springer, 2018.

[23] BENNETT, T., WANG, J. B. "Quantum optimisation via maximally amplified states", *arXiv preprint arXiv:2111.00796*, 2021.

[24] BENNETT, T., MATWIEJEW, E., MARSH, S., et al. "Quantum walk-based vehicle routing optimisation", *Frontiers in Physics*, v. 9, pp. 730856, 2021.

[25] SLATE, N., MATWIEJEW, E., MARSH, S., et al. "Quantum walk-based portfolio optimisation", *Quantum*, v. 5, pp. 513, 2021.

[26] GOLDEN, J., BÄRTSCHI, A., O'MALLEY, D., et al. "Threshold-based quantum optimization". In: *2021 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 137–147. IEEE, 2021.

[27] WANG, Z., RUBIN, N. C., DOMINY, J. M., et al. "X Y mixers: Analytical and numerical results for the quantum alternating operator ansatz", *Physical Review A*, v. 101, n. 1, pp. 012320, 2020.

[28] COOK, J., EIDENBENZ, S., BÄRTSCHI, A. "The quantum alternating operator ansatz on maximum k–vertex cover". In: *2020 IEEE International Conference on Quantum Computing and Engineering (QCE)*, pp. 83–92. IEEE, 2020.

[29] BENNETT, C. H., BERNSTEIN, E., BRASSARD, G., et al. "Strengths and weaknesses of quantum computing", *SIAM Journal on Computing*, v. 26, n. 5, pp. 1510–1523, 1997.

[30] HEADLEY, D., WILHELM, F. K. "Problem-size-independent angles for a Grover-driven quantum approximate optimization algorithm", *Physical Review A*, v. 107, n. 1, pp. 012412, 2023.

[31] HOEL, P. G., PORT, S. C., STONE, C. J. *Introduction to probability theory.* 1 ed. Boston, MA, Houghtion Mifflin, 1971.

[32] PISHRO-NIK, H. *Introduction to probability, statistics and random processes.* Kappa Research, LLC, 2014.

[33] HEADLEY, D. K. *Angles and devices for quantum approximate optimization.* Ph.D. thesis, Universität des Saarlandes, Saarbrücken, Saarland, Germany, 2023. DOI:10.22028/D291-41113.

[34] BRIDI, G. A., MARQUEZINO, F. D. L. "Analytical results for the Quantum Alternating Operator Ansatz with Grover Mixer", *arXiv preprint arXiv:2401.11056*, 2024.

[35] ZALKA, C. “Grover’s quantum searching algorithm is optimal”, *Physical Review A*, v. 60, n. 4, pp. 2746, 1999.

[36] HAMANN, A., DUNJKO, V., WÖLK, S. “Quantum-accessible reinforcement learning beyond strictly epochal environments”, *Quantum Machine Intelligence*, v. 3, pp. 1–18, 2021.

[37] DURR, C., HOYER, P. “A quantum algorithm for finding the minimum”, *arXiv preprint quant-ph/9607014*, 1996.

[38] BULGER, D., BARITOMPA, W. P., WOOD, G. R. “Implementing pure adaptive search with Grover’s quantum algorithm”, *Journal of Optimization Theory and Applications*, v. 116, pp. 517–529, 2003.

[39] BARITOMPA, W. P., BULGER, D. W., WOOD, G. R. “Grover’s quantum algorithm applied to global optimization”, *SIAM Journal on Optimization*, v. 15, n. 4, pp. 1170–1184, 2005.

[40] GILLIAM, A., WOERNER, S., GONCIULEA, C. “Grover adaptive search for constrained polynomial binary optimization”, *Quantum*, v. 5, pp. 428, 2021.

[41] ZHANG, Z., PAREDES, R., SUNDAR, B., et al. “Grover-QAOA for 3-SAT: Quadratic Speedup, Fair-Sampling, and Parameter Clustering”, *arXiv preprint arXiv:2402.02585*, 2024.

[42] VAN ROSSUM, G., DRAKE, F. L. *The Python Language Reference Manual.* Network Theory Ltd., 2011.

[43] HUNTER, J. D. “Matplotlib: A 2D graphics environment”, *Computing in Science & Engineering*, v. 9, n. 3, pp. 90–95, 2007. doi: 10.1109/MCSE.2007.55.

[44] MONAGAN, M. B., GEDDES, K. O., HEAL, K. M., et al. *Maple V Programming Guide: For Release 5.* Springer Science & Business Media, 2012.

[45] WOLFRAM, S. *The mathematica book.* 5 ed. Champaign, IL, Wolfram Research, Inc., 2003.

[46] SPANOS, A. *Probability theory and statistical inference: Empirical modeling with observational data.* 1 ed. Cambridge, Cambridge University Press, 1999.

[47] SUGIYAMA, M. *Introduction to statistical machine learning.* Burlington, MA, Morgan Kaufmann, 2015.

[48] HOEL, P. G., PORT, S. C., STONE, C. J. *Introduction to statistical theory.* 1 ed. Boston, Houghtion Mifflin, 1971.

[49] DEBNATH, L., BASU, K. "A short history of probability theory and its applications", *International Journal of Mathematical Education in Science and Technology*, v. 46, n. 1, pp. 13–39, 2015.

[50] DEFFNER, S., CAMPBELL, S. *Quantum Thermodynamics: An introduction to the thermodynamics of quantum information.* San Rafael, CA, Morgan & Claypool Publishers, 2019.

[51] LINDEBERG, T. "Scale-space for discrete signals", *IEEE Transactions on Pattern Analysis and Machine intelligence*, v. 12, n. 3, pp. 234–254, 1990.

[52] LINDEBERG, T. *Scale-space theory in computer vision*, v. 256. 1 ed. Dordrecht, Kluwer Academic Publishers, 1994.

[53] ABRAMOWITZ, M., STEGUN, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, v. 55. Washington, DC, US Government Printing Office, 1948.

[54] MICHALOS, A. C. *Encyclopedia of quality of life and well-being research.* Dordrecht, Springer Netherlands, 2014.

[55] VON HIPPEL, P. T. "Mean, median, and skew: Correcting a textbook rule", *Journal of Statistics Education*, v. 13, n. 2, 2005.

[56] WESTFALL, P. H. "Kurtosis as peakedness, 1905–2014. RIP", *The American Statistician*, v. 68, n. 3, pp. 191–195, 2014.

[57] OPPENHEIM, A. V., WILLSKY, A. S., NAWAB, S. H., et al. *Signals and systems*, v. 2. 2 ed. Upper Saddle River, NJ, Prentice-Hall, 1997.

[58] KAYE, P., LAFLAMME, R., MOSCA, M. *An introduction to quantum computing.* 1 ed. New York, Oxford University Press, 2007.

[59] MARQUEZINO, F. D. L., PORTUGAL, R., LAVOR, C. *A primer on quantum computing.* 1 ed. Cham, Springer, 2019.

[60] PORTUGAL, R. "Basic quantum algorithms", *arXiv preprint arXiv:2201.10574*, 2022.

[61] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., et al. *Introduction to algorithms.* 3 ed. Cambridge, MIT Press, 2009.

[62] SZWARCFITER, J. L. *Teoria computacional de grafos: Os Algoritmos.* 1 ed. Rio de Janeiro, Elsevier Brasil, 2018.

[63] SKIENA, S. S. *The algorithm design manual*, v. 2. London, Springer-Verlag, 2008.

[64] BONDY, J. A., MURTY, U. S. R. *Graph theory with applications*, v. 290. 1 ed. London, Macmillan, 1976.

[65] COOK, S. A. "The complexity of theorem–proving procedures". In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, p. 151–158. Association for Computing Machinery, 1971.

[66] KARP, R. M. "Reducibility among Combinatorial Problems". In: Miller, R. E., Thatcher, J. W., Bohlinger, J. D. (Eds.), *Complexity of Computer Computations*, pp. 85–103, Boston, MA, Springer US, 1972.

[67] KANN, V. *On the approximability of NP-complete optimization problems.* Ph.D. thesis, Royal Institute of Technology Stockholm, Stockholm, Sweden, 1992.

[68] GOEMANS, M. X., WILLIAMSON, D. P. "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming", *Journal of the ACM (JACM)*, v. 42, n. 6, pp. 1115–1145, 1995.

[69] ARORA, S., LUND, C., MOTWANI, R., et al. "Proof verification and the hardness of approximation problems", *Journal of the ACM (JACM)*, v. 45, n. 3, pp. 501–555, 1998.

[70] HÅSTAD, J. "Some optimal inapproximability results", *Journal of the ACM (JACM)*, v. 48, n. 4, pp. 798–859, 2001.

[71] LASRY, G. *A methodology for the cryptanalysis of classical ciphers with search metaheuristics.* Kassel, Kassel University Press GmbH, 2018.

[72] WOLSEY, L. A. *Integer programming.* 2 ed. New York, John Wiley & Sons, 2020.

[73] ABBAS, A., AMBAINIS, A., AUGUSTINO, B., et al. "Quantum optimization: Potential, challenges, and the path forward", *arXiv preprint arXiv:2312.02279*, 2023.

[74] JIANG, Z., RIEFFEL, E. G., WANG, Z. "Near-optimal quantum circuit for Grover's unstructured search using a transverse field", *Physical Review A*, v. 95, n. 6, pp. 062317, 2017.

[75] MORALES, M. E., TLYACHEV, T., BIAMONTE, J. "Variational learning of Grover's quantum search algorithm", *Physical Review A*, v. 98, n. 6, pp. 062333, 2018.

[76] BOYER, M., BRASSARD, G., HØYER, P., et al. "Tight bounds on quantum searching", *Fortschritte der Physik: Progress of Physics*, v. 46, n. 4-5, pp. 493–505, 1998.

[77] KITAEV, A. Y. "Quantum measurements and the Abelian stabilizer problem", *arXiv preprint quant-ph/9511026*, 1995.

[78] BRASSARD, G., HOYER, P., MOSCA, M., et al. "Quantum amplitude amplification and estimation", *Contemporary Mathematics*, v. 305, pp. 53–74, 2002.

[79] HARROW, A. W., HASSIDIM, A., LLOYD, S. "Quantum algorithm for linear systems of equations", *Physical Review Letters*, v. 103, n. 15, pp. 150502, 2009.

[80] AMBAINIS, A. "Variable time amplitude amplification and quantum algorithms for linear algebra problems". In: *STACS'12 (29th Symposium on Theoretical Aspects of Computer Science)*, v. 14, pp. 636–647. LIPIcs, 2012.

[81] BENTLEY, J. L., YAO, A. C.-C. "An almost optimal algorithm for unbounded searching", *Information Processing Letters*, v. 5, pp. 82–87, 1976.

[82] BAEZA-YATES, R., SALINGER, A. "Fast intersection algorithms for sorted sequences", *Algorithms and Applications: Essays Dedicated to Esko Ukkonen on the Occasion of His 60th Birthday*, pp. 45–61, 2010.

[83] BULGER, D. W., WOOD, G. R. "Hesitant adaptive search for global optimisation", *Mathematical Programming*, v. 81, pp. 89–102, 1998.

[84] GERSHENFELD, N. A. *The nature of mathematical modeling.* 1 ed. Cambriage, Cambridge University Press, 1999.

[85] ANAND, A., SCHLEICH, P., ALPERIN-LEA, S., et al. "A quantum computing view on unitary coupled cluster theory", *Chemical Society Reviews*, v. 51, n. 5, pp. 1659–1684, 2022.

[86] GRIMSLEY, H. R., ECONOMOU, S. E., BARNES, E., et al. "An adaptive variational algorithm for exact molecular simulations on a quantum computer", *Nature Communications*, v. 10, n. 1, pp. 3007, 2019.

[87] KANDALA, A., MEZZACAPO, A., TEMME, K., et al. "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets", *Nature*, v. 549, n. 7671, pp. 242–246, 2017.

[88] BITTEL, L., KLIESCH, M. "Training variational quantum algorithms is NP-hard", *Physical Review Letters*, v. 127, n. 12, pp. 120502, 2021.

[89] MCCLEAN, J. R., BOIXO, S., SMELYANSKIY, V. N., et al. "Barren plateaus in quantum neural network training landscapes", *Nature Communications*, v. 9, n. 1, pp. 4812, 2018.

[90] KINGMA, D. P., BA, J. "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980*, 2014.

[91] KÜBLER, J. M., ARRASMITH, A., CINCIO, L., et al. "An adaptive optimizer for measurement-frugal variational algorithms", *Quantum*, v. 4, pp. 263, 2020.

[92] SPALL, J. C. "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation", *IEEE Transactions on Automatic Control*, v. 37, n. 3, pp. 332–341, 1992.

[93] NAKANISHI, K. M., FUJII, K., TODO, S. "Sequential minimal optimization for quantum-classical hybrid algorithms", *Physical Review Research*, v. 2, n. 4, pp. 043158, 2020.

[94] PARRISH, R. M., IOSUE, J. T., OZAETA, A., et al. "A Jacobi diagonalization and Anderson acceleration algorithm for variational quantum algorithm parameter optimization", *arXiv preprint arXiv:1904.03206*, 2019.

[95] ARRASMITH, A., CEREZO, M., CZARNIK, P., et al. "Effect of barren plateaus on gradient-free optimization", *Quantum*, v. 5, pp. 558, 2021.

[96] WANG, S., FONTANA, E., CEREZO, M., et al. "Noise-induced barren plateaus in variational quantum algorithms", *Nature Communications*, v. 12, n. 1, pp. 6961, 2021.

[97] ZHOU, L., WANG, S.-T., CHOI, S., et al. "Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices", *Physical Review X*, v. 10, n. 2, pp. 021067, 2020.

[98] BIAMONTE, J. "Universal variational quantum computation", *Physical Review A*, v. 103, n. 3, pp. L030401, 2021.

[99] PERUZZO, A., MCCLEAN, J., SHADBOLT, P., et al. “A variational eigenvalue solver on a photonic quantum processor”, *Nature Communications*, v. 5, n. 1, pp. 4213, 2014.

[100] MCCLEAN, J. R., ROMERO, J., BABBUSH, R., et al. “The theory of variational hybrid quantum-classical algorithms”, *New Journal of Physics*, v. 18, n. 2, pp. 023023, 2016.

[101] ANSCHUETZ, E., OLSON, J., ASPURU-GUZIK, A., et al. “Variational quantum factoring”. In: *Quantum Technology and Optimization Problems: First International Workshop, QTOP 2019, Munich, Germany, March 18, 2019, Proceedings 1*, pp. 74–85. Springer, 2019.

[102] BRAVO-PRIETO, C., LAROSE, R., CEREZO, M., et al. “Variational quantum linear solver”, *Quantum*, v. 7, pp. 1188, 2023.

[103] SHOR, P. W. “Algorithms for quantum computation: discrete logarithms and factoring”. In: *Proceedings 35th annual symposium on foundations of computer science*, pp. 124–134. IEEE, 1994.

[104] SHOR, P. W. “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”, *SIAM Review*, v. 41, n. 2, pp. 303–332, 1999.

[105] ALBASH, T., LIDAR, D. A. “Adiabatic quantum computation”, *Reviews of Modern Physics*, v. 90, n. 1, pp. 015002, 2018.

[106] BORN, M., FOCK, V. “Beweis des adiabatensatzes”, *Zeitschrift für Physik*, v. 51, n. 3-4, pp. 165–180, 1928.

[107] KLUBER, G. “Trotterization in Quantum Theory”, *arXiv preprint arXiv:2310.13296*, 2023.

[108] WECKER, D., HASTINGS, M. B., TROYER, M. “Training a quantum optimizer”, *Physical Review A*, v. 94, n. 2, pp. 022309, 2016.

[109] CHILDS, A. M. *Quantum information processing in continuous time*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, 2004.

[110] NELDER, J. A., MEAD, R. “A simplex method for function minimization”, *The Computer journal*, v. 7, n. 4, pp. 308–313, 1965.

[111] FLETCHER, R. *Practical methods of optimization*. 2 ed. New York, John Wiley & Sons, 1987.

[112] WANG, Z., HADFIELD, S., JIANG, Z., et al. “Quantum approximate optimization algorithm for MaxCut: A fermionic view”, *Physical Review A*, v. 97, n. 2, pp. 022304, 2018.

[113] FARHI, E., GOLDSTONE, J., GUTMANN, S., et al. “The quantum approximate optimization algorithm and the Sherrington–Kirkpatrick model at infinite size”, *Quantum*, v. 6, pp. 759, 2022.

[114] VIJENDRAN, V., DAS, A., KOH, D. E., et al. “An expressive ansatz for low-depth quantum approximate optimisation”, *Quantum Science and Technology*, v. 9, n. 2, pp. 025010, feb 2024.

[115] WURTZ, J., LOVE, P. “MaxCut quantum approximate optimization algorithm performance guarantees for p $>$ 1”, *Physical Review A*, v. 103, n. 4, pp. 042612, 2021.

[116] CAHA, L., KLIESCH, A., KOENIG, R. “Twisted hybrid algorithms for combinatorial optimization”, *Quantum Science and Technology*, v. 7, n. 4, pp. 045013, 2022.

[117] SUNDAR, B., PAREDES, R., DAMANIK, D. T., et al. “A quantum algorithm to count weighted ground states of classical spin Hamiltonians”, *arXiv preprint arXiv:1908.01745*, 2019.

[118] ROSS, S. M. *Introduction to probability models.* 9 ed. New York, Academic Press, 2007.

[119] GROSS, J. L., YELLEN, J. *Handbook of graph theory.* 1 ed. Boca Raton, FL, CRC Press, 2003.

[120] LOEHR, N. *Bijective combinatorics.* 1 ed. Boca Raton, FL, CRC Press, 2011.

[121] MCCLEAN, J. R., HARRIGAN, M. P., MOHSENI, M., et al. “Low–depth mechanisms for quantum optimization”, *PRX Quantum*, v. 2, n. 3, pp. 030312, 2021.

[122] BENCHASATTABUSE, N., BÄRTSCHI, A., GARCÍA-PINTOS, L. P., et al. “Lower Bounds on Number of QAOA Rounds Required for Guaranteed Approximation Ratios”, *arXiv preprint arXiv:2308.15442*, 2023.

[123] LAURI, J., SCAPELLATO, R. *Topics in graph automorphisms and reconstruction*, v. 432. 2 ed. Cambridge, Cambridge University Press, 2016.

[124] RIVLIN, T. J. *Chebyshev polynomials.* 1 ed. New York, John Wiley & Sons, 1974.

[125] VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python", *Nature Methods*, v. 17, n. 3, pp. 261–272, 2020.

[126] ARNOLD, B. C., BALAKRISHNAN, N., NAGARAJA, H. N. *A first course in order statistics.* 1 ed. New York, John Wiley & Sons, 1992.

[127] BLOM, G. *Statistical estimates and transformed beta–variables.* New York, John Wiley & Sons, 1958.

[128] ARLINGHAUS, S. *Practical handbook of curve fitting.* 1 ed. Boca Raton, FL, CRC press, 1994.

[129] O'HAGAN, A., LEONARD, T. "Bayes estimation subject to uncertainty about parameter constraints", *Biometrika*, v. 63, n. 1, pp. 201–203, 1976.

[130] ASHOUR, S. K., ABDEL-HAMEED, M. A. "Approximate skew normal distribution", *Journal of Advanced Research*, v. 1, n. 4, pp. 341–350, 2010.

[131] MUDHOLKAR, G. S., HUTSON, A. D. "The epsilon–skew–normal distribution for analyzing near-normal data", *Journal of Statistical Planning and Inference*, v. 83, n. 2, pp. 291–309, 2000.

Appendix A

Low-convergence Regime Gives the Maximum Amplification on Grover's Search

To prove that the low-convergence regime gives the maximum amplification on Grover's search, we show that for a fixed r, η is a strictly decreasing function on ρ. Then, we need to prove that $\frac{d\eta}{d\rho} < 0$ for $0 < \rho \leq \sin^2(\pi/(4r+2))$. We define $R = 2r+1$ (constant) and $u = R\arcsin(\sqrt{\rho})$. By a simple substitution, the equivalent interval of u in terms of ρ is $0 < u \leq \pi/2$. Thus,

$$\eta = \frac{\sin^2(u)}{\sin^2(u/R)}. \tag{A.1}$$

Taking the derivative of η with respect to ρ, by chain rule on u, gives

$$\frac{d\eta}{d\rho} = \frac{d\eta}{du}\frac{du}{d\rho}. \tag{A.2}$$

Since

$$\frac{du}{d\rho} = \frac{R}{2\sqrt{\rho(1-\rho)}} \tag{A.3}$$

is positive and u is a one-to-one correspondence of ρ on the equivalent interval, we can analyze the derivative $\frac{d\eta}{du}$ directly with the variable u. Furthermore, we can ignore the squared on sine functions of Eq. (A.1) because if $\sin(u)/\sin(u/R)$—which is positive—is strictly decreasing on the whole interval, η also does. Therefore, taking the derivative of $\sin(u)/\sin(u/R)$ gives

$$\frac{d\sqrt{\eta}}{du} = \frac{\cos(u)\sin(u/R) - \cos(u/R)\sin(u)/R}{\sin^2(u/R)}. \tag{A.4}$$

To prove that Eq. (A.4) is negative on $0 < u \leq \pi/2$, since the denominator $\sin^2(u/R)$ is positive, we need to prove that the numerator is negative. To get that, we consider the strong condition that the numerator is strictly decreasing and tends to 0 on $u \to 0$. Firstly, the limit on $u \to 0$ is 0 from a simple substitution. Then, to show the claimed monotonicity, we take the derivative of the numerator, which is

$$\left(\frac{1}{R^2} - 1\right) \sin(u/R) \sin(u). \tag{A.5}$$

As $R \geq 3$, Eq. (A.5) is negative, as desired.

Appendix B

Proof of Theorem 14

Firstly, note that if $F_Y(T) > \rho_{Th}(r)$, by Eq. (4.70), $E_r(t)$ is monotonically non-decreasing and therefore since $P(\rho, r)$ is continuous, it is enough to prove that the monotonicity change at most one time on $F_Y(T) \leq \rho_{Th}(r)$ interval. To get it, using $\rho = F_Y(T)$ to simplify notation, by Eq. (4.63) and $\eta = P(\rho, r)/\rho$,

$$E_r(t) = \mu - \frac{G_Y(T)}{1-\rho} + G_Y(T)\frac{\eta}{1-\rho}. \tag{B.1}$$

Taking derivative of $E_r(t)$ with respect to T gives

$$\begin{aligned}\frac{dE_r(t)}{dT} &= -\frac{f_X^G(T)T(1-\rho) + f_X^G(T)G_Y(T)}{(1-\rho)^2} + f_X^G(T)T\frac{\eta}{1-\rho} \\ &+ f_X^G(T)G_Y(T)\frac{\eta_o(1-\rho)+\eta}{(1-\rho)^2} \\ &= f_X^G(T)\left(-\frac{T(1-\rho)+G_Y(T)}{(1-\rho)^2} + T\frac{\eta}{1-\rho} + G_Y(T)\frac{\eta_o(1-\rho)+\eta}{(1-\rho)^2}\right),\end{aligned} \tag{B.2}$$

where $\frac{d\eta}{dT} = f_X^G(T)\eta_o$. Explicitly,

$$\begin{aligned}\eta_o &= \frac{\sin\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right)}{\rho^2} \\ &\left(\frac{(2r+1)\sqrt{\rho}\cos\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right)}{\sqrt{1-\rho}} - \sin\left((2r+1)\arcsin\left(\sqrt{\rho}\right)\right)\right).\end{aligned} \tag{B.3}$$

By definition, $f_X^G(T)$ is non-negative (it is 0 on the points that do not belong to R_X), and then, as we are dealing with non-increasing/non-decreasing monotonicity, we can ignore it. Moreover, we can ignore points in which $\rho = 0$ as candidates of minimum since its expectation value is μ. As we are interested in the sign of

Eq. (B.2), we can multiply it by positive factors. Then, set $R = 2r + 1$ and

$$D_1 = \frac{\sqrt{\rho(1-\rho)}}{R \arcsin\left(\sqrt{\rho}\right)}. \tag{B.4}$$

Multiplying the expression by the convenient positive factor $\frac{(1-\rho)D_1}{\eta}$, we get

$$T\left(1 - \frac{1}{\eta}\right) D_1 - G_Y(T)\frac{1}{\eta}\frac{D_1}{1-\rho} + G_Y(T)\left(\frac{\eta_o}{\eta} + \frac{1}{1-\rho}\right) D_1. \tag{B.5}$$

Denoting

$$D = \left(\frac{\eta_o}{\eta} + \frac{1}{1-\rho}\right) D_1, \ D_2 = \frac{D_1}{1-\rho}, \tag{B.6}$$

we can rewrite Eq. (B.5) as

$$T\left(1 - \frac{1}{\eta}\right) D_1 - G_Y(T)\frac{1}{\eta} D_2 + G_Y(T) D. \tag{B.7}$$

At $T \to -\infty$, the derivative begins negative, which follows from $\eta > 1$, $G_Y(T) = 0$, and $D_1 > 0$. Furthermore, we demonstrate further that D is negative. That way, if $T \geq 0$, the expression does not change the sign since all terms are non-negative. So, we assume $T < 0$. As $G_Y(T)$ is non-increasing, η is strictly decreasing, and $\eta > 1$, if we prove that (i) D_1 is strictly decreasing, (ii) D_2 is strictly increasing, and (iii) D is strictly decreasing and negative, all terms of Eq. (B.7) are non-decreasing and the monotonicity of $E_r(t)$ change one time on this interval, proving the theorem. The minimum of the original discrete function is hit either on $\rho \leq \rho_{Th}(r)$ or in the smallest defined $F_Y(T)$ in which $\rho > \rho_{Th}(r)$.

Consider the substitution $u = R \arcsin\left(\sqrt{\rho}\right)$. By the same argument used for η on Appendix A, we can directly analyze the derivative with respect to u on the interval $0 < u \leq \pi/2$. Thus,

$$D_1 = \frac{\sin\left(2u/R\right)}{2u}, \ D_2 = \frac{\tan\left(u/R\right)}{u}, \tag{B.8}$$

and as,

$$\begin{aligned} \frac{\eta_o}{\eta} D_1 &= \frac{R\sqrt{\rho}\cos\left(R\arcsin\left(\sqrt{\rho}\right)\right) - \sqrt{1-\rho}\sin\left(R\arcsin\left(\sqrt{\rho}\right)\right)}{R\arcsin\left(\sqrt{\rho}\right)\sqrt{\rho}\sin\left(R\arcsin\left(\sqrt{\rho}\right)\right)} \\ &= \frac{R\sin\left(u/R\right)\cos\left(u\right) - \cos\left(u/R\right)\sin\left(u\right)}{u\sin\left(u/R\right)\sin\left(u\right)} = \frac{R\cot\left(u\right) - \cot\left(u/R\right)}{u}, \end{aligned} \tag{B.9}$$

then

$$D = \frac{R\cot\left(u\right) - \cot\left(u/R\right)}{u} + \frac{\tan\left(u/R\right)}{u}. \tag{B.10}$$

We deal with (i), (ii) with a similar argument as done for η. Taking the derivatives of both

$$\frac{dD_1}{du} = \frac{2u\cos(2u/R) - R\sin(2u/R)}{2Ru^2}, \quad \frac{dD_2}{du} = \frac{u\sec^2(u/R) - R\tan(u/R)}{Ru^2}. \tag{B.11}$$

Both denominators are positive for $u > 0$, and the limit of numerators on $u \to 0$ is 0. Then, taking the derivatives of the numerators,

$$-\frac{4u\sin(2u/R)}{R}, \quad \frac{2u\sec^2(u/R)\tan(u/R)}{R}, \tag{B.12}$$

for (i) and (ii) cases, respectively. As claimed, the first is negative and the last positive on the whole u interval.

The claimed *(iii)* is more complicated. First, we must prove that the limit with $u \to 0$ is negative. For that, consider the expansion of $\cot(x)$ with the Taylor series $\sin(x) = x - x^3/6 + \mathcal{O}(x^5)$ as $x \to 0$ and $\cos(x) = 1 - x^2/2 + \mathcal{O}(x^4)$ as $x \to 0$,

$$\begin{aligned} \cot(x) &= \cos(x)\frac{1}{\sin(x)} = \left(1 - \frac{x^2}{2} + \mathcal{O}(x^4)\right)\left(\frac{1}{x - \frac{x^3}{6} + \mathcal{O}(x^5)}\right) \\ &= \frac{1}{x}\left(1 - \frac{x^2}{2} + \mathcal{O}(x^4)\right)\left(\frac{1}{1 - \frac{x^2}{6} + \mathcal{O}(x^4)}\right) \end{aligned} \tag{B.13}$$

Replacing

$$\frac{1}{1-x} = 1 + x + \mathcal{O}(x^2) \text{ as } x \to 0 \tag{B.14}$$

in Eq. (B.13) gives

$$\cot(x) = \frac{1}{x}\left(1 - \frac{x^2}{2} + \mathcal{O}(x^4)\right)\left(1 + \frac{x^2}{6} + \mathcal{O}(x^4)\right) = \frac{1}{x} - \frac{x}{3} + \mathcal{O}(x^3) \tag{B.15}$$

as $x \to 0$. Furthermore, we use the well-known expansion

$$\tan(x) = \sum_{n=1}^{\infty} \frac{B_{2n}(-4)^n(1-4^n)}{2n(2n-1)!}x^{2n-1} = x + \mathcal{O}(x^3) \tag{B.16}$$

as $x \to 0$, where B_{2n} denotes the Bernoulli number [53], that can be expressed in terms of Riemann zeta function $\zeta(s) = \sum_{j=1}^{\infty}\frac{1}{j^s}$ [53] by

$$B_{2n} = \frac{(-1)^{n+1}2(2n)!}{(2\pi)^{2n}}\zeta(2n). \tag{B.17}$$

Thus, replacing the expansions in Eq. (B.10),

$$D = \left(-\frac{R}{3} + \frac{1}{3R} + \mathcal{O}(u^2)\right) + \left(\frac{1}{R} + \mathcal{O}(u^2)\right) = \frac{4 - R^2}{3R} + \mathcal{O}(u^2), \tag{B.18}$$

and we have a negative limit

$$\lim_{u\to 0} D = \frac{4-R^2}{3R}, \tag{B.19}$$

since $R \geq 3$. That way, if we prove that the derivative of D is negative, we establish that D is both negative and strictly decreasing.

The technique used on the proof is an extension of the technique employed at η, D_1, and D_2 cases, i.e., for each derivative, we show that $u \to 0$ is non-positive and ignoring positive factors, consider the stronger condition that the next derivative is also negative.

Beginning with the derivative of D,

$$\begin{aligned}\frac{dD}{du} &= \frac{R\cot(u/R) - R^2\cot(u) + u\csc^2(u/R) - R^2u\csc^2(u)}{Ru^2} \\ &+ \frac{u\sec^2(u/R) - R\tan(u/R)}{Ru^2}.\end{aligned} \tag{B.20}$$

The denominators are positive on $u > 0$. To get the limit of $u \to 0$ on the numerator, we plug the expansions of all trigonometric functions. The expansions of $\cot(x)$ and $\tan(x)$ have already been introduced, the expansion of $\sec^2(x)$ is given by $\sec^2(x) = \frac{d\tan(x)}{dx} = 1 + \mathcal{O}(x^2)$, and $\csc^2(x)$ is given analogously to the expansion of $\cot(x)$ as

$$\begin{aligned}\csc^2(x) &= \frac{1}{\sin^2(x)} = \frac{1}{x^2 - \frac{x^4}{3} + \mathcal{O}(x^6)} = \frac{1}{x^2}\frac{1}{1-\frac{x^2}{3}+\mathcal{O}(x^4)} \\ &= \frac{1}{x^2} + \frac{1}{3} + \mathcal{O}(x^2)\end{aligned} \tag{B.21}$$

as $x \to 0$, with the expansion of $\sin^2(x)$ following from

$$\begin{aligned}\sin^2(x) &= \frac{1}{2}(1-\cos(2x)) = \frac{1}{2}\left(1 - \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n)!}(2x)^{2n}\right) \\ &= \sum_{n=1}^{\infty}\frac{(-1)^{n+1}2^{2n-1}}{(2n)!}x^{2n} = x^2 - \frac{x^4}{3} + \mathcal{O}(x^6)\end{aligned} \tag{B.22}$$

as $x \to 0$. Combining them, we conclude that the numerator has an order of $\mathcal{O}(u^3)$, and then, the limit is 0. Now, we must satisfy

$$\begin{aligned}&R\cot(u/R) - R^2\cot(u) + u\csc^2(u/R) - R^2u\csc^2(u) + u\sec^2(u/R) \\ &- R\tan(u/R) < 0,\end{aligned} \tag{B.23}$$

which gives the derivative

$$\begin{aligned}
&2uR^2\left(\frac{\cos(u)}{\sin^3(u)} - \frac{\cos(u/R)}{R^3\sin^3(u/R)} + \frac{\sin(u/R)}{R^3\cos^3(u/R)}\right)\\
&= 2uR^2\left(\frac{\cos(u)}{\sin^3(u)} + \frac{\sin^4(u/R) - \cos^4(u/R)}{R^3\sin^3(u/R)\cos^3(u/R)}\right)\\
&= 2uR^2\left(\frac{\cos(u)}{\sin^3(u)} + \frac{8(\sin^2(u/R) - \cos^2(u/R))(\sin^2(u/R) + \cos^2(u/R))}{R^3\sin^3(2u/R)}\right)\\
&= 2uR^2\left(\frac{\cos(u)}{\sin^3(u)} - \frac{8\cos(2u/R)}{R^3\sin^3(2u/R)}\right),
\end{aligned} \tag{B.24}$$

where we use the trigonometric identity $\cos(2x) = \cos^2(x) - \sin^2(x)$ in the last equality. The inequality can be manipulated as

$$\begin{aligned}
&2uR^2\left(\frac{\cos(u)}{\sin^3(u)} - \frac{8\cos(2u/R)}{R^3\sin^3(2u/R)}\right) < 0\\
\Rightarrow\ &\frac{1}{2uR^2}\left(\frac{\sin^3(u)}{\cos(u)} - \frac{R^3\sin^3(2u/R)}{8\cos(2u/R)}\right) > 0.
\end{aligned} \tag{B.25}$$

The limit with $u \to 0$ inside the parenthesis is immediately 0. Ignoring $\frac{1}{2uR^2}$ and take the derivative for the last time,

$$\tan^2(u) + 2\sin^2(u) - \frac{R^2}{4}(\tan^2(2u/R) + 2\sin^2(2u/R)) \tag{B.26}$$

and so

$$\tan^2(u) + 2\sin^2(u) > \frac{R^2}{4}(\tan^2(2u/R) + 2\sin^2(2u/R)). \tag{B.27}$$

The proof of Eq. (B.27) is based on the Taylor series expansion of $2\sin^2(x)$ and $\tan^2(x)$. The first is given by Eq. (B.22) and the second can be computed using the expansion of $\tan(u)$ as

$$\begin{aligned}
&\tan^2(x) = \sec^2(x) - 1 = \frac{d\tan(x)}{dx} - 1\\
&= \frac{d}{dx}\sum_{n=1}^{\infty}\frac{B_{2n}(-4)^n(1-4^n)}{2n(2n-1)!}x^{2n-1} - 1 = \sum_{n=1}^{\infty}\frac{B_{2n+2}(-4)^{n+1}(1-4^{n+1})}{(2n+2)(2n)!}x^{2n}\\
&= \sum_{n=1}^{\infty}\frac{(-1)^n 2(2n+2)!(-4)^{n+1}(1-4^{n+1})\zeta(2n+2)}{(2\pi)^{2n+2}(2n+2)(2n)!}x^{2n}\\
&= \sum_{n=1}^{\infty}\frac{2(2n+1)4^{n+1}(4^{n+1}-1)\zeta(2n+2)}{(2\pi)^{2n+2}}x^{2n}.
\end{aligned} \tag{B.28}$$

Note that both expansions have terms of the same order, which allows a direct comparison. We are interested in establishing that all terms of $\tan^2(x) + 2\sin^2(x)$ expansion are non-negative. The sine squared expansion alternates the sign, having a negative sign for n even, while tangent squared expansion has all positive terms. Then, is enough to show that all even n terms of $\tan^2(x)$ expansion are equal or greater than the absolute value of the respective terms of $2\sin^2(x)$. By Eq. (B.22) and (B.28), we must satisfy the inequality

$$\frac{2(2n+1)4^{n+1}(4^{n+1}-1)\zeta(2n+2)}{(2\pi)^{2n+2}} \geq \frac{2^{2n}}{(2n)!} \tag{B.29}$$

for all even n. If $n = 2$, since $\zeta(6) = \pi^6/945$, both sides of Eq. (B.29) have the same value of $2/3$. Using inequalities $\zeta(2n+2) \geq 1$ and $4^{n+1}(4^{n+1}-1) \geq \pi^{2n+2}$, the bound is reduced to

$$\frac{2(2n+1)}{2^{2n+2}} \geq \frac{2^{2n}}{(2n)!} \ \Rightarrow\ \frac{(2n+1)!}{16^n} \geq 2. \tag{B.30}$$

We take it by induction. For $n = 4$, the inequality holds since $2835/512 > 2$. Then, we assume that is true for any $n > 4$ even. The left side of the inequality can be written for $n+2$ as

$$\frac{(2(n+2)+1)!}{16^{n+2}} = \left(\frac{(2n+5)(2n+4)(2n+3)(2n+2)}{256}\right)\left(\frac{(2n+1)!}{16^n}\right). \tag{B.31}$$

The second parenthesis is the bound of n, and the first is larger than 1. Therefore, our claim follows by induction hypothesis.

To finish, we demonstrate that each expansion term on the left side of Eq. (B.27) is equal to or greater than the analog terms of the right side expansion. To get that, consider an arbitrary term n of the expansions. The argument of trigonometric functions on the left side is $x = u$ while the left side is $x = 2u/R$. Combining it with the constant multiplication factor $R^2/4$ on the right size, we must satisfy

$$u^{2n} \geq \frac{R^2}{4}\left(\frac{2u}{R}\right)^{2n} \ \Rightarrow\ 1 \geq \left(\frac{2}{R}\right)^{2n-2}. \tag{B.32}$$

They are equal for $n = 1$, and as $R \geq 3$, the left side is larger for $n > 1$. Therefore, since we prove that all non-zero expansion terms are positives, the left side is larger than the right for any $u > 0$, and the inequality of Eq. (B.27) follows, establishing the theorem.