[Tutorial] Quantum Approximate Optimization Algorithm

Thanatach Tangsakul

In this tutorial, we will examine what is a Quantum Approximate Optimization Algorithm and how is it able to find a good approximate solution to an optimization problem. An explanation will be complement with an animation of the algorithm which will be introduced when we have enough context in mind.

Quantum Approximate Optimization Algorithm (QAOA) [1] is a certain heuristic optimization algorithm that have been received considerable amount of attention over the past few years due to the potential in performance improvement over existing classical algorithms [2] and its effectiveness on NISQ devices [3]. It aims to find a good approximate solution to a combinatorial optimization problem. As the name suggests, the problem involves finding an optimal combination from every valid configuration. Of course, a valid configuration depends on the problem instance at hand. For example, it could be a string with certain number of 1s or it could even be a spanning tree of a certain graph. But the number of valid configurations must be finite. Thus, we could more concisely and precisely define the combinatorial optimization problem to be the problem of finding an optimal object from a finite set of discrete objects.

Max-Cut problem is a standard problem of this type. It involves coloring an undirected graph with two colors in order to maximize a number of special edges. An edge is special if its endpoints do not the same color. It can be seen that given a graph with n vertices, we can denote any coloring using an n-bit string. An example can be seen in Figure 1 where 0011 denotes the color of vertices $0, 1, 2, 3$ respectively and special edges are colored in red. From the graph, notice that 0011 and 1100 give the same answer. We shall denote $Cost(x)$, or shortly just $C(x)$, to be the value of a solution x from now on. So, notationally, $C(0011) = C(1100)$.

In general, for any combinatorial optimization problem, we can uniquely map any possible solution to an n-bit string, or we can say that we can map the solution space to the set of n-bit strings. Thus, in the context of QAOA, we will think of a basis state as a correspondence to one possible solution. Therefore, with an appropriate mapping of solution, the optimal solution will correspond to some basis state.

It should be noted that there might be more than one way to define a cost function for any problem at hand. Certain function might allow for a lower number of unitary operations than other functions and certain function might allow for a better amplitude amplification.

1 Overview of QAOA

Now, we are ready to look at the QAOA circuit which can be seen in Figure 2.

As in any typical quantum algorithm, $H^{\otimes n}$ prepares an equal superposition of every basis state. Other than that, the circuit consists of two types of

Figure 1: An example of a Max-Cut problem

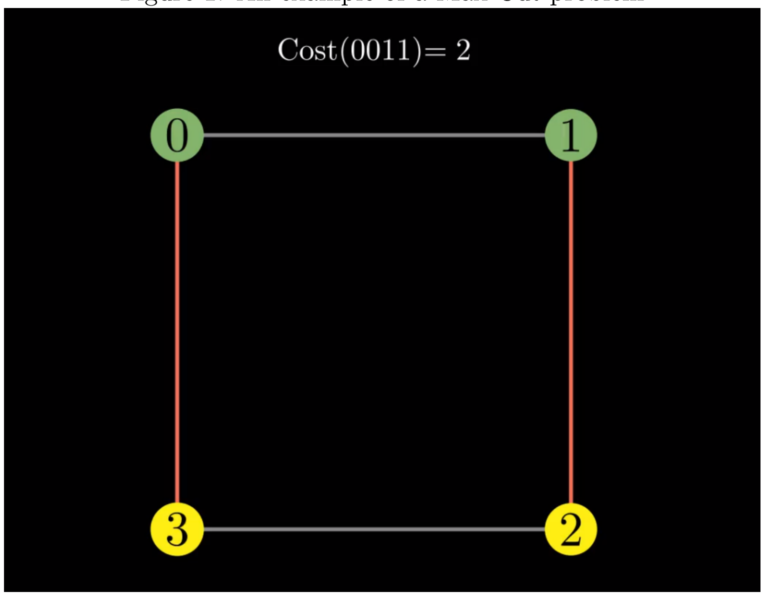
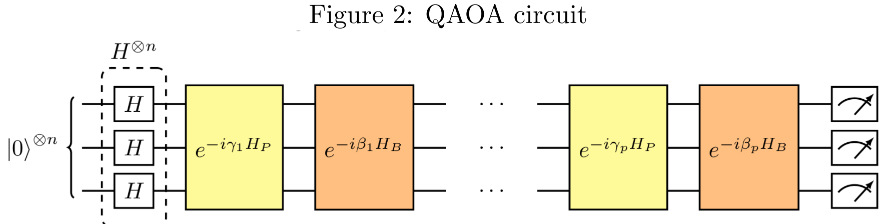


Figure 2: QAOA circuit



important operators, a phase operator $e^{-i\gamma_j H_P}$ and a mixing operator $e^{-i\beta_j H_B}$. Those two together, in the correct order, are considered to be one layer of QAOA. We can perform those two together any number of times to create any number of layers, the standard notation for the number of layers is p. So the total number of important operators is $2p$.

H_B and H_P are certain Hamiltonian which we will examine later. $\gamma_1, \beta_1, \gamma_2, \beta_2, ..., \gamma_p, \beta_p$ are certain parameters whose values have to be found in order to discover a good solution. To find those values, we employ a classical computer to train the parameters, similar to the training process in Machine Learning. By

1. Initialize the parameters with some values
2. Evaluate how good are those parameters from the measurement
3. Try varying the parameters in a way that might achieve a better result

4. Stop or go back to 2.

An exact detail of the training process depends on the employed training method of which has many to choose from [4]. However, a simple training method can help illustrate the concept which is that in the Step 3, we randomly increase or decrease the kth parameter by some small ϵ_k. If we did not achieve a better result, we will vary those parameters in the inverse direction instead.

Typically, the goodness of the parameters are defined to be the expected value of an answer from the measurement, or notationally $\sum p(x)C(x)$ where $p(x)$ is the probability of measuring $|x\rangle$. So, when the problem is about finding the minimum(maximum) $C(x)$, the lower(higher) the expected value the better. However, other definitions also exist such as the average of the best and the second-best answer receiving from measuring the circuit 1024 times.

Hence, with a high level overview, one can view QAOA as simply a variational algorithm. But we will inspect further what makes QAOA stands out.

Next, we will examine the phase operator and the mixing operator in great detail in order to obtain insights into how the states can be amplified.

2 Phase Operator $e^{-i\gamma_j H_P}$

For any combinatorial optimization problem at hand such that we can map the solution space to the set of all n-bit strings, we will define

$$H_P = \sum_{x \in \{0,1\}^n} C(x)\ |x\rangle\langle x|$$

Assuming that $C(x)$ of the problem is a real number, H_P will be a Hermitian matrix.

From the fact that if H is a Hermitian matrix with a spectral decomposition $\sum_i \lambda_i\ |\psi_i\rangle\langle\psi_i|$, $e^{i\theta H}$ will be $\sum_i e^{i\theta\lambda_i}\ |\psi_i\rangle\langle\psi_i|$. Then

$$e^{-i\gamma_j H_P} = \sum_{x \in \{0,1\}^n} e^{-i\gamma_j C(x)}\ |x\rangle\langle x|$$

which means that for any $x \in \{0,1\}^n$

$$e^{-i\gamma_j H_P}|x\rangle = e^{-i\gamma_j C(x)}|x\rangle$$

In other words, the phase of the state $|x\rangle$ will be rotated by an angle $-\gamma_j C(x)$ when applying the phase operator $e^{-i\gamma_j H_P}$ as in the animation.

There is one question raised "why do we perform this operator even if the amplitude is not change?" We can see later that this operator is important. But we can give an intuitive answer that "this is to imbue cost information which is a unique characteristic of the problem instance into the quantum state".

Moreover, there is another important question "how to apply $e^{-i\gamma_j H_P}$". The answer to this question for any Max-Cut instance or for other interesting problems can be found in the footnote [5].

3 Mixing Operator $e^{-i\beta_j H_B}$

Since analyzing the mixing operator in the same fashion as analyzing the phase operator will give us little to no insight, we will turn towards analyzing the mixing operator as a sequence of local operators. We will show that the mixing operator can be analyzed locally as an exchange of amplitude between basis states that have hamming distance equal to one.

3.1 Decomposition

H_B is defined to be $X_1 + X_2 + ... + X_n$ where X_k is $I^{\otimes(k-1)} \otimes X \otimes I^{n-k}$. Therefore,

$$\begin{aligned} e^{-i\beta_j H_B} &= e^{-i\beta_j (X_1 + X_2 + \cdots + X_n)} \\ &= e^{-i\beta_j X_1 - i\beta_j X_2 - \ldots - i\beta_j X_n} \end{aligned} \tag{1}$$

Since X_i is commute with X_j for any $1 \leq i, j \leq n$, by induction,

$$e^{-i\beta_j H_B} = e^{-i\beta_j X_1} e^{-i\beta_j X_2} \cdots e^{-i\beta_j X_n} \tag{2}$$

Therefore, the mixing operator consists of n partial operators which we shall examine now.

3.2 Partial operator $e^{-i\beta_j X_k}$

Let us first analyze the first partial operator $e^{-i\beta_j X_1}$. By definition, $X_1 = X \otimes I^{\otimes(n-1)}$. Then $X_1^2 = I \otimes I^{\otimes(n-1)} = I^{\otimes n}$.

Since $e^{i\theta A} = \cos(\theta)I + i\sin(\theta)A$, for any $A_{N\times N}$ such that $A^2 = I_{N\times N}$. Therefore,

$$\begin{aligned} e^{-i\beta_j X_1} &= \cos(-\beta_j) I_{N\times N} + i\sin(-\beta_j) X_1 \\ &= \cos(\beta_j) I_{N\times N} - i\sin(\beta_j) X_1 \\ &= \cos(\beta_j) I \otimes I^{\otimes(n-1)} - i\sin(\beta_j) X \otimes I^{\otimes(n-1)} \\ &= [\cos(\beta_j) I - i\sin(\beta_j) X] \otimes I^{\otimes(n-1)} \end{aligned} \tag{3}$$

For simplicity, suppose that we have $|\psi\rangle = \alpha_0|0\rangle|x\rangle + \alpha_1|1\rangle|x\rangle = (\alpha_0|0\rangle + \alpha_1|1\rangle) \otimes |x\rangle$ for some $x \in \{0,1\}^{n-1}$. Then

$$\begin{aligned} e^{-i\beta_j X_1}|\psi\rangle &= [(\cos(\beta_j)I - i\sin(\beta_j)X)(\alpha_0|0\rangle + \alpha_1|1\rangle)] \otimes |x\rangle \\ &= \quad [\ \alpha_0(\cos(\beta_j)I - i\sin(\beta_j)X)|0\rangle \\ &\quad + \alpha_1(\cos(\beta_j)I - i\sin(\beta_j)X)|1\rangle\ ]\ \otimes |x\rangle \\ &= \quad [\ \alpha_0\cos(\beta_j)|0\rangle - \alpha_0 i\sin(\beta_j)|1\rangle \\ &\quad + \alpha_1\cos(\beta_j)|1\rangle - \alpha_1 i\sin(\beta_j)|0\rangle\ ]\ \otimes |x\rangle \\ &= \quad (\alpha_0\cos(\beta_j) - \alpha_1 i\sin(\beta_j))\ |0\rangle|x\rangle \\ &\quad + (\alpha_1\cos(\beta_j) - \alpha_0 i\sin(\beta_j))\ |1\rangle|x\rangle \end{aligned} \tag{4}$$

This means that if we have

$$|\psi\rangle = \sum_{x\in\{0,1\}^{n-1}} \alpha_{0x}|0\rangle|x\rangle + \alpha_{1x}|1\rangle|x\rangle$$

After applying $e^{-i\beta_j X_1}$ we get

$$\begin{aligned} |\psi'\rangle = e^{-i\beta_j X_1}|\psi\rangle = \sum_{x\in\{0,1\}^{n-1}} & \left[\alpha_{0x}\cos(\beta_j) - \alpha_{1x} i \sin(\beta_j)\right] |0\rangle|x\rangle \\ & + \left[\alpha_{1x}\cos(\beta_j) - \alpha_{0x} i \sin(\beta_j)\right] |1\rangle|x\rangle \end{aligned} \tag{5}$$

In other words, for any $x \in \{0,1\}^{n-1}$, the amplitude of the state $|0\rangle|x\rangle$ after applying $e^{-i\beta_j X_1}$ depends only on the amplitude of $|1\rangle|x\rangle$ and vice versa. For example, the new amplitude of $|0000\rangle$ will depend only on the amplitude of $|1000\rangle$. Pedagogically, we could say that they pair up and exchanges amplitude with each other according to Eq. 5. Hence, the highlighting of all vertical edges in the animation when $e^{-i\beta_j X_1}$ is being applied.

With an appropriate grouping pair of states, the analysis can also be applied to $e^{-i\beta_j X_k}$ for any $1 \le k \le n$. Therefore, when we apply any partial operator $e^{-i\beta_j X_k}$, the new amplitude of a basis state $|x\rangle$ will depend only on the basis state $X_k|x\rangle$ with the same pattern illustrated above.

3.3 Other mixing operators

It is important to emphasize that other definitions of H_B can be used instead of the current one. The current one is the mixing operator from the original paper of QAOA which is called an X-mixer. For other mixers, such as an XY-mixer [reference], they might give a better performance, compared to the X-mixer. However, in order to understand the gist of QAOA, the analysis of the X-mixer should suffice.

Animation

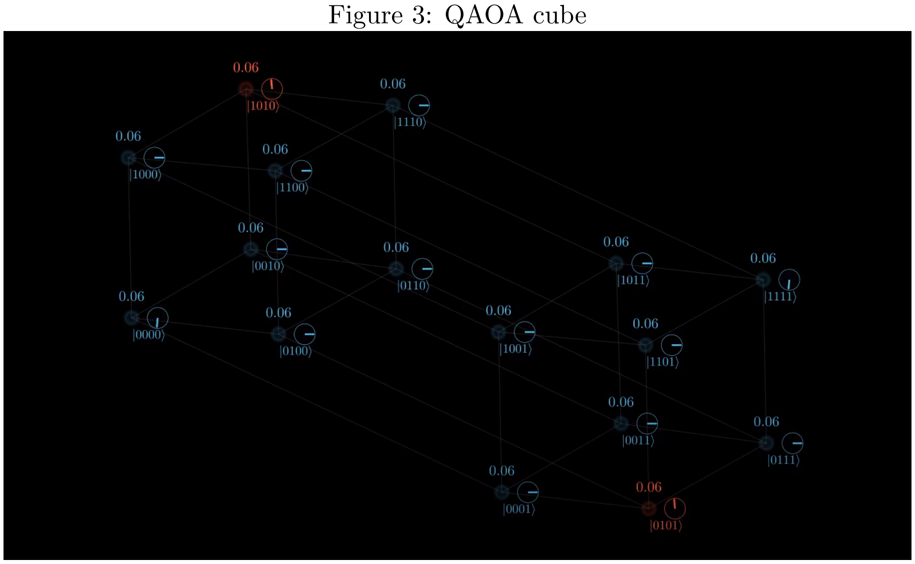
Now we should have enough context in mind to introduce an animation of the QAOA algorithm that can greatly complement the explanation. This will also be crucial when we look at various amplification behaviors that can occur in QAOA circuits. The video can be found in https://youtu.be/JXBRODOLgnQ. First, let me explain the meaning of each object in the animation. (For convenience, the picture can be found in Figure 3)

- Each sphere refers to certain basis state denoted by a nearby ket. Its color indicates the quality of the solution, i.e., red indicates that it is the best solution; otherwise, the color will be blue.
- The clock of each sphere indicates the phase of its amplitude, e.g., if the amplitude of a state $|1010\rangle = e^{-i\cdot\frac{3\pi}{2}} \cdot \frac{1}{\sqrt{16}}$, the angle of the hand will be at $-3\pi/2$ rad.

- The number above each sphere refers to the probability **round to two decimal places** that the measurement will result in that state.
- An edge will connect two spheres if their states are distinct by one Hamming distance, e.g., $|1010\rangle$ is connected to $|1110\rangle$.

In the animation, p = 2 and $[\gamma_1\ \beta_1\ \gamma_2\ \beta_2] \approx$ [-1.15 1.01 0.56 -0.43] which obtained from the training process. The problem instance that QAOA trying to solve in the animation is a Max-Cut problem with the graph shown in Figure 1. **But $C(x)$ is the number of special edges minus the number of non-special edges**. This is due to the simplicity when applying the phase operator. Please note that the animation starts with the system after $H^{\otimes n}$ has been applied. As you can see, after all the operators have been applied, the probability of measuring the state corresponding to the best solution is very close to 1. We will examine why the change of amplitude in the animation is in that way and how can we guide the training process to converge to the set of parameters that amplify the best states in the next sections.

Figure 3: QAOA cube



4 How the states get amplified

4.1 Quantify the amplitude change

Consider a complex number $z = \alpha_0 \cos(\beta) - \alpha_1 i \sin(\beta)$, where $\alpha_0 = r_0 e^{i\theta_0}$ and $\alpha_1 = r_1 e^{i\theta_1}$ for some real number θ_0, θ_1, $r_0 \neq 0$, $r_1 \neq 0$. And let denote $\Delta\theta = \theta_0 - \theta_1$, $k = \frac{r_1}{r_0}$, $p_0 = r_0^2$, and $p_0' = |z|^2$. Note that z has the same form as the amplitude of $|0\rangle|x\rangle$ in Eq. 5.

First by showing that,

$$\frac{p_0'}{p_0} \;=\; 1 \;+\; (k^2-1)\sin^2(\beta) \;+\; k\sin(2\beta)\sin(-\Delta\theta)$$

This comes from

$$\begin{aligned}
|z|^2 &= |\alpha_0|^2\cos^2(\beta) + |\alpha_1|^2\sin^2(\beta) + 2\Re(\alpha_0\bar{\alpha_1}\cos(\beta)\sin(\beta)i) \\
&= r_0^2\cos^2(\beta) + r_1^2\sin^2(\beta) + r_0r_1\sin(2\beta)\Re(e^{i\Delta\theta}i) \\
&= r_0^2\cos^2(\beta) + r_1^2\sin^2(\beta) - r_0r_1\sin(2\beta)\sin(\Delta\theta) \\
&= r_0^2\cos^2(\beta) + r_1^2\sin^2(\beta) + r_0r_1\sin(2\beta)\sin(-\Delta\theta)
\end{aligned} \tag{6}$$

And

$$\begin{aligned}
\frac{p_0'}{p_0} = \frac{|z|^2}{r_0^2} &= \cos^2(\beta) + k^2\sin^2(\beta) + k\sin(2\beta)\sin(-\Delta\theta) \\
&= \cos^2(\beta) + \sin^2(\beta) - \sin^2(\beta) + k^2\sin^2(\beta) + k\sin(2\beta)\sin(-\Delta\theta) \\
&= 1 + (k^2-1)\sin^2(\beta) + k\sin(2\beta)\sin(-\Delta\theta)
\end{aligned} \tag{7}$$

This concludes the first part. As we expect, the proportionality depends on both β and $\Delta\theta$. One little fact we should note here is that when $\beta = 0$, $\frac{p_0'}{p_0}$ will be equal to 1 which means that the amplitude does not change. This can also be seen from the fact that $e^{-i\cdot 0\cdot X_k}$ is identical to an identity operator.

Next, before jumping into calculus, let us examine further what happens when $k = 1$ and $\beta = \pi/4$.

$$\begin{aligned}
\frac{p_0'}{p_0} \;&=\; 1 \;+\; (1^2-1)\sin^2(\pi/4) \;+\; 1\cdot\sin(2\cdot\pi/4)\sin(-\Delta\theta) \\
&=\; 1 \;+\; \sin(-\Delta\theta)
\end{aligned} \tag{8}$$

This means that the ratio depends hugely on the $\Delta\theta$. Observe that when $\Delta\theta = -\frac{\pi}{2}$, the ratio is at its maximum which means that the basis state with former amplitude α_0 (such as the state $|0\rangle|x\rangle$ when applying $e^{-i\beta_j X_1}$) will be amplified after applying a single partial operator. Notice that the amplitude of a basis state with former amplitude α_1 (such as the state $|1\rangle|x\rangle$ when applying $e^{-i\beta_j X_1}$) will be accordingly reduced, which we can quantify its ratio too; however, we can obtain interesting results by analyzing only the ratio $\frac{p_0'}{p_0}$.

4.2 Analyze the role of the β parameter

Let us examine the derivative of the equation $\frac{p_0'}{p_0}$. For notational convenience, let us define the function

$$f(\beta, k, \Delta\theta) = 1 \;+\; (k^2-1)\sin^2(\beta) \;+\; k\sin(2\beta)\sin(-\Delta\theta)$$

Since $f(\beta, k, \Delta\theta) = f(\beta+\pi, k, \Delta\theta)$, we can restrict ourselves to analyze only when $\beta \in [-\pi/2, \pi/2]$. For a similar reason we can analyze only for $\Delta\theta \in [-\pi, \pi]$.

It follows that

$$
\begin{aligned}
f'(\beta, k, \Delta\theta) = \frac{\partial f}{\partial \beta} &= 2(k^2 - 1)\sin(\beta)\cos(\beta) + 2k\cos(2\beta)\sin(-\Delta\theta) \\
&= (k^2 - 1)\sin(2\beta) + 2k\cos(2\beta)\sin(-\Delta\theta)
\end{aligned} \tag{9}
$$

Let us analyze case by case before make any further inference. There are five cases in total:

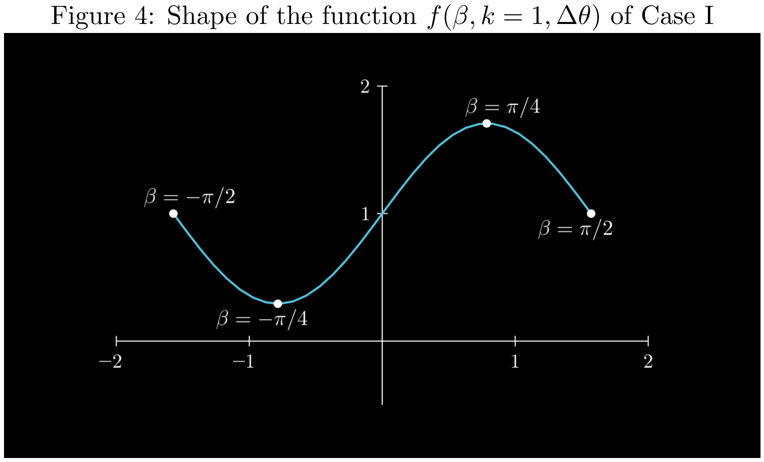
Case I: $k = 1$ and $0 < -\Delta\theta < \pi$,

$$f'(\beta, k, \Delta\theta) = 2\cos(2\beta)\sin(-\Delta\theta)$$

i) if $\beta \in (-\pi/4, \pi/4)$, $f' > 0$
ii) if β is $-\pi/4$ or $\pi/4$, $f' = 0$
iii) if $\beta \in (-\pi/2, \pi/4)$ or $\beta \in (\pi/4, \pi/2)$, $f' < 0$

Together with the fact that $f(0, k, \Delta\theta) = 0$, the shape of the function f can be visualized as in Figure 4.

Figure 4: Shape of the function $f(\beta, k = 1, \Delta\theta)$ of Case I



Case II: $k > 1$ and $0 < -\Delta\theta < \pi$,

As before, we want to analyze the function f' to infer some important facts about f

i) $-\pi/4 < \beta < \pi/4$

First, if β, k, and $\Delta\theta$ are real numbers satisfy the constraints Case II.i such that

$$f'(\beta, k, \Delta\theta) = (k^2 - 1)\sin(2\beta) + 2k\cos(2\beta)\sin(-\Delta\theta) = 0$$

Because k is not 1, it implies that

$$\sin(2\beta) = -\frac{2k}{k^2-1}\cos(2\beta)sin(-\Delta\theta)$$

Since $\cos(2\beta)$ is not 0,

$$\tan(2\beta) = -\frac{2k}{k^2-1}sin(-\Delta\theta) \tag{10}$$

If $k = 1+\epsilon$ for some small ϵ, it implies that β must be $-\pi/4+\delta$ for some small δ. (We will denote ϵ and δ to be some small real number greater than 0)

It also follows that if $k = 1+\epsilon$ for some small ϵ, there will exist some turning point $\beta' = -\pi/4+\delta$ for some small δ such that the Eq. 10 is satisfied which implies that f' will be 0. (Let us call a β' to be a "turning point" if it makes $f' = 0$)

Because for any $k \neq 1$

$$\frac{d\frac{-k}{k^2-1}}{dk}(k) = \frac{-(k^2-1)-(-k)(2k)}{(k^2-1)^2} = \frac{k^2+1}{(k^2-1)^2} > 0 \tag{11}$$

and $\tan(2\beta)$ is an increasing function over the domain of $\beta \in (-\pi/4, \pi/4)$, it implies that, for a fixed $\Delta\theta$, when k is increased, the turning point will be increased too. And when $k \to \infty$, the turning point will approach 0.

Similarly, if $\beta, k, \Delta\theta$ are real numbers satisfy the constraints Case II.i such that

$$f'(\beta, k, \Delta\theta) = (k^2-1)\sin(2\beta) + 2k\cos(2\beta)\sin(-\Delta\theta) > 0$$

Because k is greater than 1, it implies that

$$\sin(2\beta) > -\frac{2k}{k^2-1}\cos(2\beta)sin(-\Delta\theta)$$

Since $\cos(2\beta)$ is greater than 0,

$$\tan(2\beta) > -\frac{2k}{k^2-1}sin(-\Delta\theta) \tag{12}$$

Because $\tan(2\beta)$ is an increasing function over the domain of $\beta \in (-\pi/4, \pi/4)$, β will satisfy this if and only if it is greater than the turning point.

Before we move on, you can glimpse at the shape of the function f when $-\pi/4 < \beta < \pi/4$ shown in Figure 5 and Figure 6. As you can see, when k is increased, the turning point shifts towards 0.

ii) $\pi/4 < \beta \leq \pi/2$ or $-\pi/2 \leq \beta < -\pi/4$

We will proceed in the same fashion. First, it can be shown with the same argument as in the previous case that $f' = 0$ if and only if Eq. 10 holds. Therefore, if $k = 1 + \epsilon$, the turning point will be $\pi/4 + \delta$ for some small δ.

Since $\tan(2\beta)$ is an increasing function over the domain of β in our consideration, it implies that, for a fixed $\Delta\theta$, when k is increased, the turning point will be increased too. And when $k \to \infty$, the turning point will approach $\pi/2$.

Secondly, similar to the derivation for Eq. 12, since $\cos(2\beta)$ is less than 0, f' will be greater than 0 if and only if $\tan(2\beta) < -\frac{2k}{k^2-1} sin(-\Delta\theta)$.

So, only when $\beta \in (\pi/4, \beta')$, where β' is the turning point, that can make $f' > 0$.

iii) $\beta = \pi/4$ or $\beta = -\pi/4$

When $\beta = \pi/4$, $f' > 0$.

When $\beta = -\pi/4$, $f' < 0$.

We have accounted for every subcase. Together with the fact that when β is 0, f will be 1, the shape of the function f can be visualized as in Figure 5 and Figure 6.

One important thing to note is that when $k = 1$, the maximum point is exactly at $\beta = \pi/4$ and the minimum point is at $-\pi/4$. And when k is increased, the maximum point and the minimum point will shift towards the right.

We probably expect that when k is decreased from 1, the maximum point and the minimum point will shift towards the left. We will show this in the next case which will proceed in the similar fashion.

Figure 5: Shape of the function $f(\beta, k = 1.25, \Delta\theta)$ of Case II

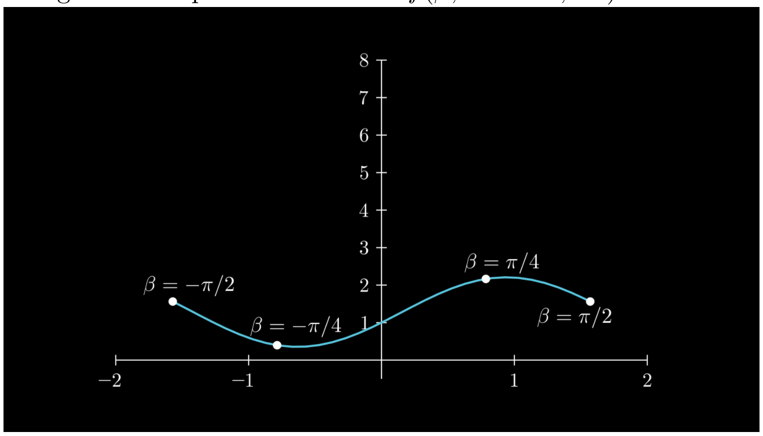
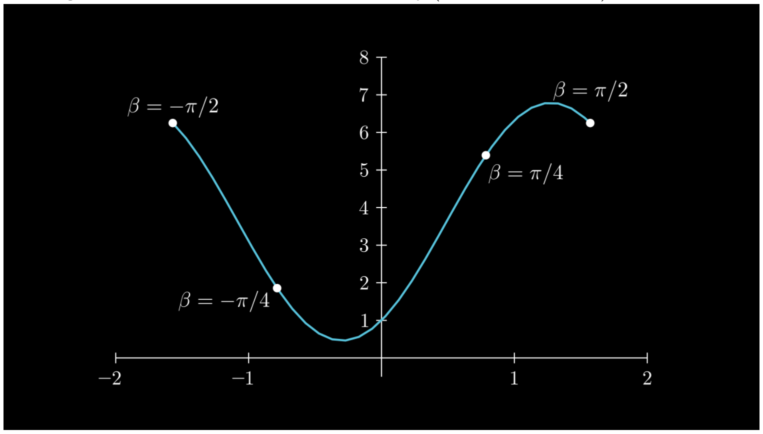


Figure 6: Shape of the function $f(\beta, k = 2.5, \Delta\theta)$ of Case II



Case III: $k < 1$ and $0 < -\Delta\theta < \pi$,

i) $-\pi/4 < \beta < \pi/4$

As before, it can be shown that f' will be 0 if and only if

$$\tan(2\beta) = -\frac{2k}{k^2 - 1} sin(-\Delta\theta)$$

as in the Eq. 10. This implies that if $k = 1 - \epsilon$ for some small ϵ, the turning point β' will be $\pi/4 - \delta$ for some small δ. Since $\frac{-k}{k^2-1}$ is an increasing function over the set of $k < 1$ as shown in Eq. 11, when k is decreased, the turning point will be decreased too. And when $k = 0 + \epsilon$ for some small ϵ, the turning point will be about $0 + \delta$ for some small δ.

Similarly, f' will be greater than 0 if and only if $\tan(2\beta) < -\frac{2k}{k^2-1} sin(-\Delta\theta)$. Therefore, only when $\beta \in (-\pi/4, \beta')$, where β' is the turning point, that can make $f' > 0$.

ii) $\pi/4 < \beta \leq \pi/2$ or $-\pi/2 \leq \beta < -\pi/4$

Again, with the same derivation in Eq. 10, if $k = 1 - \epsilon$ for some small ϵ, the turning point will be $-\pi/4 - \epsilon$ for some small δ. And when k is decreased, the turning point will be decreased too. When $k = 0 + \epsilon$ for some small ϵ, the turning point will be $-\pi/2 + \delta$ for some small δ.

With a similar derivation as in Eq. 12, except $k < 1$ and $\cos(2\beta) < 0$, f' will be greater than 0 if and only if $\tan(2\beta) > -\frac{2k}{k^2-1} sin(-\Delta\theta)$. Thus, only when $\beta \in (\beta', -\pi/4)$ where β' is the turning point.

iii) $\beta = \pi/4$ or $\beta = -\pi/4$

When $\beta = \pi/4$, $f' < 0$.

When $\beta = -\pi/4$, $f' > 0$.

Now, we have analyzed every subcase. The shape of the function f can be summarized as in Figure 7 and Figure 8. As we have already guessed, when k is decreased, the maximum point shifts to the left and the minimum point shifts to the right.

Figure 7: Shape of the function $f(\beta, k = 0.75, \Delta\theta)$ of Case III

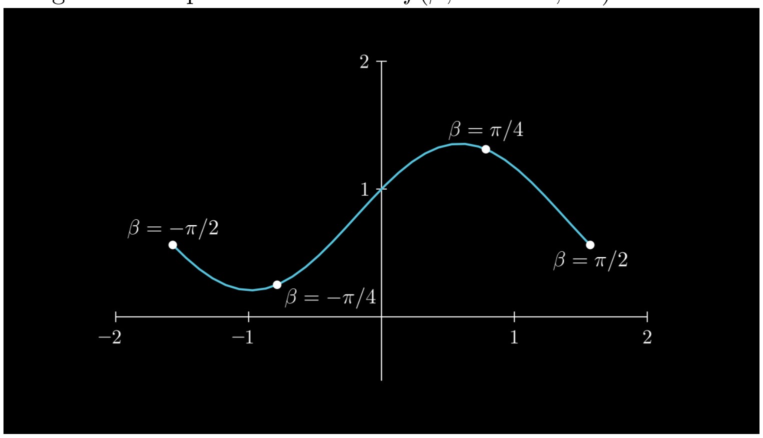
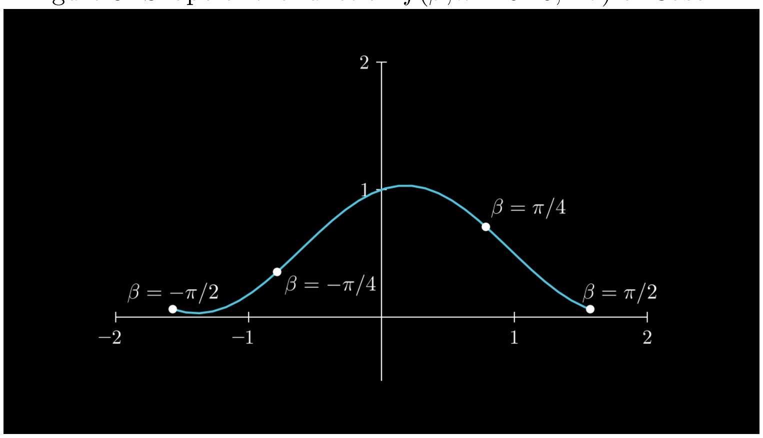


Figure 8: Shape of the function $f(\beta, k = 0.25, \Delta\theta)$ of Case III



Case IV: $-\Delta\theta = 0$ or π

This case is relative simple, compared to other cases.

i) if $k = 1$, $f = 1$
ii) if $k > 1$, $f \geq 1$, and f is at its maximum when β is $\pi/2$ or $-\pi/2$
iii) if $k < 1$, $f \leq 1$, and f is at its maximum when β is 0

Case V: $0 < \Delta\theta < \pi$

Since $f(\beta, k, \Delta\theta) = f(-\beta, k, -\Delta\theta)$, it implies that

i) $k = 1$

$f(\beta, 1, \Delta\theta)$ will be equal to $f(-\beta, 1, -\Delta\theta)$ which falls into the Case I. Therefore, the shape of the function will be the same as in Figure 4 but flip around the y-axis.

ii) $k > 1$

Similarly, for some fixed k and $\Delta\theta$, the function f will be the same as the function $f(\beta, k, -\Delta\theta)$ but flip around the y-axis. So the shape of the function will be as the function that falls in Case II but flip around the y-axis.

iii) $k < 1$

With the same reasoning, the shape of the function will be as the function that falls in Case III but flip around the y-axis.

Concrete Example

Suppose we are applying the operator $e^{-i\beta_j X_1}$, and consider a pair of $|0\rangle|x\rangle$ and $|1\rangle|x\rangle$ for some $x \in \{0,1\}^{n-1}$ such that the parameters fall into Case I or Case II or Case III.

If $|0\rangle|x\rangle$ is better than $|1\rangle|x\rangle$, we would want to amplify the state $|0\rangle|x\rangle$. Suppose that we are applying this at the first layer. k will be equal to 1 because every state has the same amplitude. To maximize the amplification, we would want β to be $\pi/4$.

If $|1\rangle|x\rangle$ is better than $|0\rangle|x\rangle$, and $k = 4$, which means the shape of f will be closely similar to Figure 6. We would want to amplify the state $|1\rangle|x\rangle$ (which is equivalent to decrease the magnitude of the state $|0\rangle|x\rangle$), so our optimal β would be far from $-\pi/4$ and close to 0.

It should be emphasized that for any partial operator, the effect of β can be analyzed similarly.

4.3 Combining partial operators

So far, we have analyzed only the partial operator. To complete the picture, we have to take into account the effect of applying n partial operators. This is rather complicated to analyze due to the phase change of each state after applying a single partial operator. Moreover, with only one governing parameter β, some partial operators might amplify the best state whereas others might reduce its amplitude. (This leads us to introduce more governing parameters per layer such as having n parameters, each governs a single partial operator [6].)

Despite the complexity, for the best state to be amplified, it must be amplified in at least one partial operator. Since we know the condition of how a state can be amplified when applying a single partial operator, we have useful insights into what are the necessary conditions of the value of a parameter that will be able to amplify the best state as a whole mixing operator, and we can practically use this insight to choose a more educated initial set of parameters for a training procedure so that it can converge to a good set of parameters.

5 QAOA's Amplification

Now, we are ready to talk about how QAOA is able to amplify the best solution.

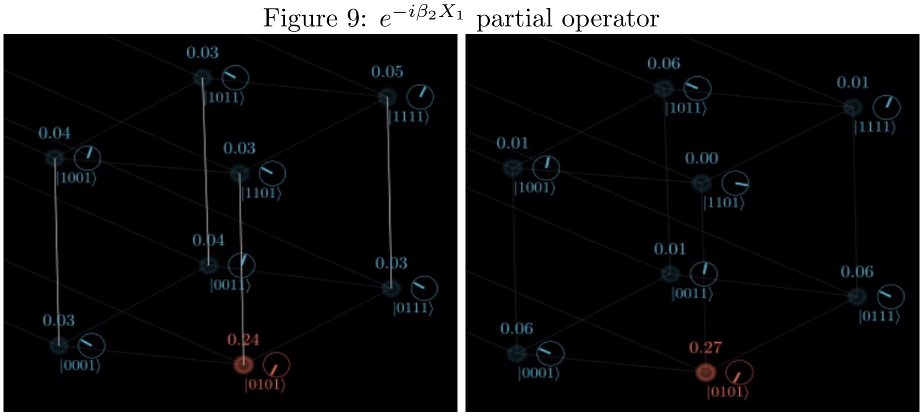
5.1 Accumulation

In this situation, the amplitude of the best state is increased cumulatively after applying each partial operator. When this type of amplification occurs, the β_j parameter governing the layer usually have the property that $|\beta_j| \leq \pi/4$. This can be well-explained by looking at the animation's second layer operator.

Example

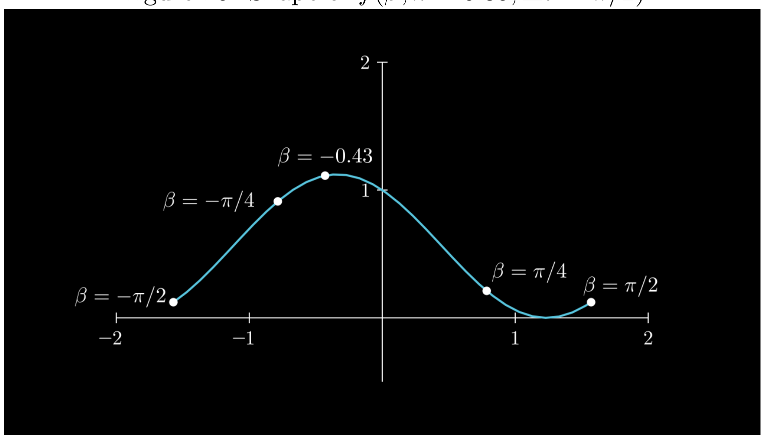
Let us focus on the state $|0101\rangle$ and $|1101\rangle$ while the operator $e^{-i\beta_2 X_1}$ with $\beta_2 \approx -0.43$ is being applied. This is shown in Figure 9.

Figure 9: $e^{-i\beta_2 X_1}$ partial operator



We can see that $\Delta\theta \approx \pi/2$ and $k \approx \frac{\sqrt{0.03}}{\sqrt{0.24}} \approx 0.35 < 1$. This falls into Case V which is the same as Case III but the shape of f will flip around the y-axis. So the shape can be more precisely visualized as in Figure 10. From the shape

Figure 10: Shape of $f(\beta, k = 0.35, \Delta\theta = \pi/2)$



of the function, we can immediately distinguish a value of the β parameter that can amplify $|0101\rangle$. As we can see, β_2 is close to the optimal parameter that will amplify $|0101\rangle$. Therefore, $|0101\rangle$ will accumulate the amplitude from $|1101\rangle$. A similar reasoning can be applied to show that $|1010\rangle$ will accumulate the amplitude from $|0010\rangle$.

And when the operator $e^{-i\beta_2 X_2}$ is applied, $|0101\rangle$ will accumulate the amplitude from $|0001\rangle$, and $|1010\rangle$ will accumulate from $|1110\rangle$ with a similar reasoning. And so on for the remaining partial operators.

Note that the amplification is possible because the phase of each state is rotated to a proper angle.

Practical guide: How to cumulatively amplify a state that its amplitude is larger than its neighbors

Consider for any appropriate grouping pair of states $|x\rangle$ and $|y\rangle$ when performing some partial operator, such as $|0001\rangle$ and $|0101\rangle$ when the operator $e^{-i\beta_2 X_2}$ is being applied. Let us suppose that $\Delta\theta$ is not 0 or π. (When it is the case, it is almost certain that $C(x) = C(y)$ and $k = 1$, which means that any amplification will not occur.) There will be a β with $|\beta| < \pi/4$ that enable us to amplify the larger state. (Whether β will be greater or less than 0 will depend on the value of $\Delta\theta$.) And the bigger the larger state, the smaller $|\beta'|$ where β' is the best parameter that amplify the larger state. However, we do not know what that value is. But we can set the initial parameter β to be in that range, so we have a better chance that the training parameter will converge to the optimal value. Because we expect that the amplitude of the best state should be increased after each mixing operator $e^{-i\beta_j H_B}$ is applied, for any $1 \leq d \leq p-1$, we should set β_d and β_{d+1} such that $|\beta_d| > |\beta_{d+1}|$, and, of course, for any $1 \leq j \leq p$, $|\beta_j| \leq \pi/4$.

For example, for $p = 3$, One possible choice for initial β might be $(\beta_1, \beta_2, \beta_3) = (\frac{\pi}{4}, \frac{\pi}{4} - 0.1, \frac{\pi}{4} - 0.2)$, and expect that the training process will find $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ that rotate the phase of each state to an appropriate angle. It should be emphasized that it is not guaranteed that the final parameters obtained from the classical training procedure will have this pattern. However, this kind of descending pattern of initial parameters has been empirically observed to be promising [7].

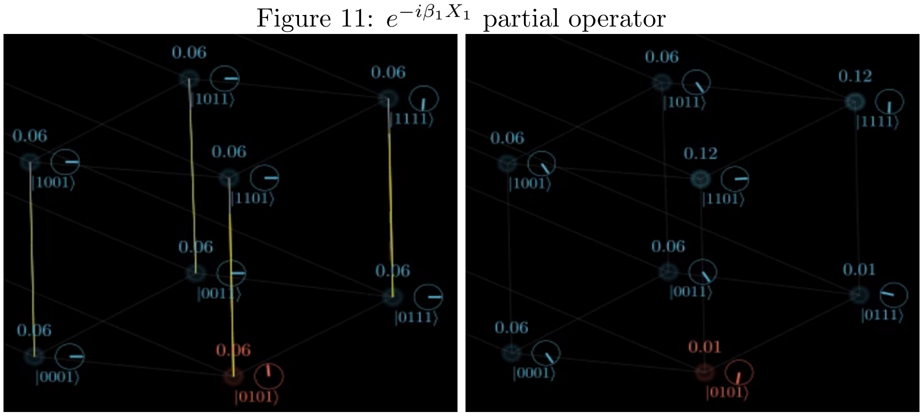
5.2 Send out

In the special situation where the best solution and the second-best solution are around the same value, this type of amplification could occur and usually occurs with $|\beta_j| \geq \pi/4$. Let us look at the animation's first layer operators as an example.

Example

Again, let us focus on the state $|0101\rangle$ and $|1101\rangle$ while the operator $e^{-i\beta_1 X_1}$ with $\beta_1 \approx 1.01$ is being applied. This is shown in Figure 11.

Figure 11: $e^{-i\beta_1 X_1}$ partial operator



We can see that $\Delta\theta \approx \pi/2$ and $k = 1$. This falls into Case V which is the same as Case I but the shape of f will flip around the y-axis. Thus, $|0101\rangle$'s amplitude will be decreased. It is important to remember that most of the amplitude of $|0101\rangle$ is currently with $|1101\rangle$.

Next, when the operator $e^{-i\beta_1 X_2}$ is being applied, with the same reasoning, most of the amplitude of $|1101\rangle$ will be with $|1001\rangle$. While $|0101\rangle$ gain a small amount of amplitude from $|0001\rangle$.

Afterwards, when $e^{-i\beta_1 X_3}$ is being applied, most of the $|1001\rangle$'s amplitude will be transferred to $|1011\rangle$.

Finally, when $e^{-i\beta_1 X_4}$ is being applied, $|1011\rangle$'s amplitude will be transferred to $|1010\rangle$ which is another best state. $|0101\rangle$ will also receive considerable amount of amplitude from $|0100\rangle$ which is accumulated from the path $|1010\rangle \to |0010\rangle \to |0110\rangle \to |0100\rangle \to |0101\rangle$ with a similar reasoning.

Practical guide: How to guide the training process to find the parameters with this type of amplification

Analogous to the previous guide, to guide the training procedure to have this amplification pattern, for any $1 \le d \le p-1$, we should set β_d and β_{d+1} such that $|\beta_d| < |\beta_{d+1}|$, and for any $1 \le j \le p$, $|\beta_j| \ge \pi/4$.

For example, for $p = 3$, One possible choice for initial β might be $(\beta_1, \beta_2, \beta_3) = (\frac{\pi}{4}, \frac{\pi}{4} + 0.1, \frac{\pi}{4} + 0.2)$, Again, it should be emphasized that it is not guaranteed that the final parameters obtained from the classical training procedure will have this pattern. However, in max cut instances, this type of amplification tend to occur due to the symmetry in its solution space, i.e., $H_P|x\rangle = H_P X^{\otimes n}|x\rangle$ for any $x \in \{0,1\}^n$. Therefore, guiding the training procedure in this fashion should allow for a better use of existing structure.

6 Important Note

Finally, some important remarks should be noted

- Amplification pattern is not always intuitive as one may think. For example, It could be the case that our final parameters end up amplifying non-optimal states up to $p-1$ layers where each layer behaves like 5.1, and at the last layer behaves similar to 5.2.
- With different underlying problem structure, certain amplification pattern may be better than others. (In terms of the minimum number of layers requires to achieve a good result, the achieved expected value, or the ease of converging to good parameters.)
- It can be seen that $e^{-i\beta_j X_{k_1}}$ and $e^{-i\beta_j X_{k_2}}$ can commute with each other. This commutative property implies that we can consider the effect of applying the partial operators in any arbitrary order. In the animation, this implies that even with different orders of operators, the amplification behavior is still the same.
- A lot of QAOA variations has also been explored. For example, A version of QAOA where the mixing operator is generalized [8] and GM-QAOA [9] where the idea of Grover's Algorithm is incorporated.
- It is important to emphasize that the goal of QAOA is not to find the best solution but rather a good approximate solution, one that gives a sufficiently good solution. However, the approximation ratio of QAOA for general p is hard to analyze. Nonetheless, the empirical observation of the performance of QAOA has been examined [10].

- Last but not least, from the analysis we have done, we can see that the QAOA's operators have a clear goal, i.e., the phase operators set up the angle between each state and the mixing operators try to amplify the best state. However, there are certain question that we have not tried to answer "how do we come up with this circuit" because they seem extremely arbitrary. The answer to this lies on the inspiration that QAOA draw from which is the quantum adiabatic algorithm. More information can be found in the original paper of QAOA [1].

References

[1] The original paper can be found here: https://arxiv.org/abs/1411.4028

[2] The paper examining the QAOA's performance can be found here: https://arxiv.org/abs/1812.01041

[3] The paper examining the effects of noise on QAOA can be found here: https://arxiv.org/abs/1909.02196

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

[5] https://qiskit.org/textbook/ch-applications/qaoa.html#Appendix

[6] https://arxiv.org/abs/2010.06660

[7] https://arxiv.org/abs/1812.01041

[8] https://arxiv.org/abs/1709.03489

[9] https://arxiv.org/abs/2006.00354

[10] https://arxiv.org/abs/1812.01041