Global Optimum Search in Quantum Deep Learning

Lanston Hau Man Chu, Tejas Bhojraj, Rui Huang

Project of CS 880: Quantum Computing

Instructor: Dieter van Melkebeek

May 8, 2020

Overview

Gradient Descent

Gradient Descent

- Gradient descent (GD) is an essential algorithm widely used in machine learning to optimize the objective function

Gradient Descent

- Gradient descent (GD) is an essential algorithm widely used in machine learning to optimize the objective function
- Key idea: Move θ iteratively towards the direction of the steepest gradient to reach the optimum

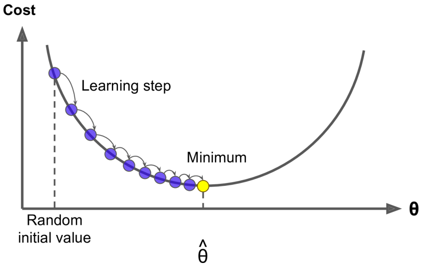
Gradient Descent

- Gradient descent (GD) is an essential algorithm widely used in machine learning to optimize the objective function
- Key idea: Move θ iteratively towards the direction of the steepest gradient to reach the optimum

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathcal{L}(\theta)$$

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

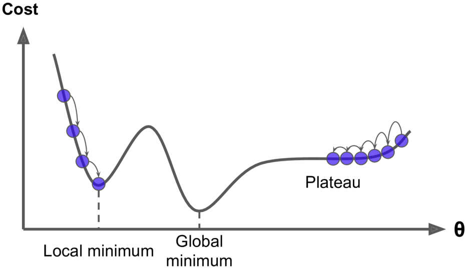
$$\theta_{k+1} = \theta_k - \eta \frac{\partial \mathcal{L}_\theta}{\partial \theta}\bigg|_{\theta=\theta_k}$$



Drawbacks of Gradient Descent

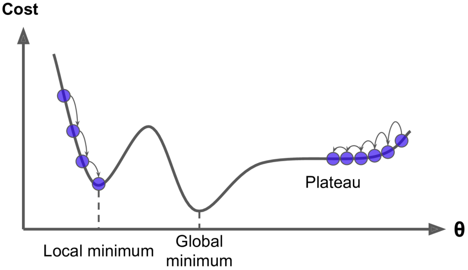
Drawbacks of Gradient Descent

- Depending on the initial point, GD may find a local optimum instead of the global optimum.



Drawbacks of Gradient Descent

- Depending on the initial point, GD may find a local optimum instead of the global optimum.



- The upper bound or expected number of iterations to reach convergence is difficult to determine.

Key Contributions

Key Contributions

- Two quantum approaches (Average Approach and PSTC Approach) to find the global optimum (instead of a local optimum) for optimizing machine learning models.

Key Contributions

- Two quantum approaches (Average Approach and PSTC Approach) to find the global optimum (instead of a local optimum) for optimizing machine learning models.
- Theoretical analyses to show that the expected cost for both approaches are $O(\sqrt{|\Theta|}N)$.

Key Contributions

- Two quantum approaches (Average Approach and PSTC Approach) to find the global optimum (instead of a local optimum) for optimizing machine learning models.
- Theoretical analyses to show that the expected cost for both approaches are $O(\sqrt{|\Theta|}N)$.
- Novel objective function maximizing the number of "cut-off indicators $\mathbb{E}_{\theta j}$" to fit the property of quantum computing in optimizing machine learning models

Key Contributions

- Two quantum approaches (Average Approach and PSTC Approach) to find the global optimum (instead of a local optimum) for optimizing machine learning models.
- Theoretical analyses to show that the expected cost for both approaches are $O(\sqrt{|\Theta|}N)$.
- Novel objective function maximizing the number of "cut-off indicators $\mathbb{E}_{\theta j}$" to fit the property of quantum computing in optimizing machine learning models
- Potential for PSTC to reduce the cost further to $O(\sqrt{|\Theta|} \cdot sublinear(N))$ in future work

Overview of Average Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta, x_j)\rangle$$

$$\theta^*_{avg} = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}_\theta^{avg} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{N}\sum_{j=1}^{N}\ell(\theta, x_j)$$

Overview of Average Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta, x_j)\rangle$$

$$\theta^*_{avg} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathcal{L}_\theta^{avg} = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \frac{1}{N}\sum_{j=1}^{N}\ell(\theta, x_j)$$

- Treat the machine learning model as a quantum blackbox

Overview of Average Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta, x_j)\rangle$$

$$\theta^*_{avg} = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}^{avg}_{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{N}\sum_{j=1}^{N}\ell(\theta, x_j)$$

- Treat the machine learning model as a quantum blackbox
- Utilize Durr & Hoyer (DH) algorithm to find the global minimum

Overview of Average Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta, x_j)\rangle$$

$$\theta^*_{avg} = \operatorname*{argmin}_{\theta\in\Theta} \mathcal{L}_\theta^{avg} = \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{N}\sum_{j=1}^{N}\ell(\theta, x_j)$$

- Treat the machine learning model as a quantum blackbox
- Utilize Durr & Hoyer (DH) algorithm to find the global minimum
- Expected number of steps: $O(N\sqrt{|\Theta|})$

Overview of Partial Swap Test Cut-off (PSTC) Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle$$

$$\theta^*_{\mathsf{PSTC}} = \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_\theta^{\mathsf{PSTC}} = \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$$

Overview of Partial Swap Test Cut-off (PSTC) Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle$$

$$\theta^*_{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}_\theta^{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$$

- The cut-off indicator $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$.

Overview of Partial Swap Test Cut-off (PSTC) Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle$$

$$\theta^*_{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}_\theta^{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$$

- The cut-off indicator $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$.
- Employ a "partial" swap test and amplitude amplification to boost the probability of finding a good θ

Overview of Partial Swap Test Cut-off (PSTC) Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle$$

$$\theta^*_{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\,\mathcal{L}_\theta^{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\,\frac{1}{N}\sum_{j=1}^{N}\mathbb{1}[\,\ell(\theta,x_j) \leq \ell_{threshold}]$$

- The cut-off indicator $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta,x_j) \leq \ell_{threshold}]$.
- Employ a “partial” swap test and amplitude amplification to boost the probability of finding a good θ
- Still $O(N\sqrt{|\Theta|})$, but only at checking process

Overview of Partial Swap Test Cut-off (PSTC) Approach

$$|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Quantum parallelism with cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle$$

$$\theta^*_{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}_\theta^{\mathsf{PSTC}} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$$

- The cut-off indicator $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$.
- Employ a "partial" swap test and amplitude amplification to boost the probability of finding a good θ
- Still $O(N\sqrt{|\Theta|})$, but only at checking process
- Enable future improvement to maybe $O(sublinear(N) \cdot \sqrt{|\Theta|})$

Overview

Approach 1: Average Approach

1. Consider the function $f(i) = f(\theta_i) := \sum_j \ell(\theta_i, x_j)$
2. Use the DH algorithm [2] to find the minimum of the set $\{f(i)\}_i$ with a slight modification: When the DH algorithm queries $f(i)$ for some i (i.e. when the algorithm makes a query on $|i\rangle|0^t\rangle$ and needs $|i\rangle|f(i)\rangle$ as output), do the following subroutine:
   - Classically compute $f(i)$

Approach 1: Average Approach

1. Consider the function $f(i) = f(\theta_i) := \sum_j \ell(\theta_i, x_j)$
2. Use the DH algorithm [2] to find the minimum of the set $\{f(i)\}_i$ with a slight modification: When the DH algorithm queries $f(i)$ for some i (i.e. when the algorithm makes a query on $|i\rangle|0^t\rangle$ and needs $|i\rangle|f(i)\rangle$ as output), do the following subroutine:
    - Classically compute $f(i)$
    - Build a unitary operator, U mapping $|i\rangle|0^t\rangle \mapsto |i\rangle|f(i)\rangle$.

Approach 1: Average Approach

1. Consider the function $f(i) = f(\theta_i) := \sum_j \ell(\theta_i, x_j)$
2. Use the DH algorithm [2] to find the minimum of the set $\{f(i)\}_i$ with a slight modification: When the DH algorithm queries $f(i)$ for some i (i.e. when the algorithm makes a query on $|i\rangle|0^t\rangle$ and needs $|i\rangle|f(i)\rangle$ as output), do the following subroutine:
   - Classically compute $f(i)$
   - Build a unitary operator, U mapping $|i\rangle|0^t\rangle \mapsto |i\rangle|f(i)\rangle$.
   - Return $U(|i\rangle|0^t\rangle)$ to the DH algorithm.

Analysis

- Each time we invoke the subroutine, we need N classical queries to ℓ.

Analysis

- Each time we invoke the subroutine, we need N classical queries to ℓ.
- The DH algorithm makes an expected number of $O(\sqrt{|\Theta|})$ queries. So, we run the subroutine $O(\sqrt{|\Theta|})$ times on average.

Analysis

- Each time we invoke the subroutine, we need N classical queries to ℓ.
- The DH algorithm makes an expected number of $O(\sqrt{|\Theta|})$ queries. So, we run the subroutine $O(\sqrt{|\Theta|})$ times on average.
- So, the cost is $O(\sqrt{|\Theta|}N)$.

Analysis

- Each time we invoke the subroutine, we need N classical queries to ℓ.
- The DH algorithm makes an expected number of $O(\sqrt{|\Theta|})$ queries. So, we run the subroutine $O(\sqrt{|\Theta|})$ times on average.
- So, the cost is $O(\sqrt{|\Theta|}N)$.
- Also, we know that the DH algorithm succeeds with probability 0.5 and so by running it many times, we get the parameter minimizing the loss with arbitrarily good accuracy at the cost $O(\sqrt{|\Theta|}N)$.

Overview

Motivation: Cost reduction

Motivation: Cost reduction

- Average approach: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N} \ell(\theta, x_j)\rangle$

Motivation: Cost reduction

- Average approach: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N} \ell(\theta, x_j)\rangle$
- Want some function f: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} \sum_{j=1}^{N} |\theta\rangle|j\rangle|f(\theta, x_j)\rangle$

Motivation: Cost reduction

- Average approach: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta, x_j)\rangle$
- Want some function f: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} \sum_{j=1}^{N}|\theta\rangle|j\rangle|f(\theta, x_j)\rangle$
- **Cut-off indicator** $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$ (1-qubit)

$$\boxed{|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle}$$

Motivation: Cost reduction

- Average approach: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N}\ell(\theta,x_j)\rangle$
- Want some function f: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} \sum_{j=1}^{N}|\theta\rangle|j\rangle|f(\theta,x_j)\rangle$
- **Cut-off indicator** $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta,x_j) \leq \ell_{threshold}]$ (1-qubit)

$$\boxed{|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle}$$

- Easy summation (by inner product):

$$\Big(\sum_{j=1}^{N}\langle j|\langle E_{\theta j}|\Big)\Big(\sum_{k=1}^{N}|k\rangle|1\rangle\Big) = \sum_{j=1}^{N}\langle j|j\rangle\langle E_{\theta j}|1\rangle = \sum_{j=1}^{N} E_{\theta j}$$

Motivation: Cost reduction

- Average approach: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(N)} |\theta\rangle|j\rangle|\sum_{j=1}^{N} \ell(\theta, x_j)\rangle$
- Want some function f: $\sum_{j=1}^{N} |\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} \sum_{j=1}^{N} |\theta\rangle|j\rangle|f(\theta, x_j)\rangle$
- **Cut-off indicator** $E_{\theta j} \triangleq \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$ (1-qubit)

$$\boxed{|\theta\rangle|j\rangle|\mathbf{0}\rangle \xmapsto{\text{Cost } O(1)} |\theta\rangle|j\rangle|E_{\theta j}\rangle}$$

- Easy summation (by inner product):

$$\Big(\sum_{j=1}^{N} \langle j|\langle E_{\theta j}|\Big)\Big(\sum_{k=1}^{N} |k\rangle|1\rangle\Big) = \sum_{j=1}^{N} \langle j|j\rangle\langle E_{\theta j}|1\rangle = \sum_{j=1}^{N} E_{\theta j}$$

- New objective function (want larger $\mathcal{L}_\theta^{\mathsf{PSTC}}$):

$$\theta^*_{\mathsf{PSTC}} = \operatorname*{argmax}_{\theta \in \Theta} \mathcal{L}_\theta^{\mathsf{PSTC}} = \operatorname*{argmax}_{\theta \in \Theta} \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}[\,\ell(\theta, x_j) \leq \ell_{threshold}]$$

Partial Swap Test ($Q_{\text{1-query}}$)

Partial Swap Test ($Q_{\text{1-query}}$)

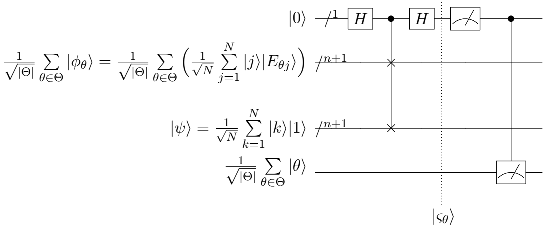
- Want inner product: $\left(\sum_{j=1}^{N}\langle j|\langle E_{\theta j}|\right)\left(\sum_{k=1}^{N}|k\rangle|1\rangle\right) = \sum_{j=1}^{N} E_{\theta j}$

$$\text{Define } \begin{cases} |\phi_\theta\rangle \triangleq \frac{1}{\sqrt{N}} \sum_{j=1}^{N} |j\rangle|E_{\theta j}\rangle \\ |\psi\rangle \triangleq \frac{1}{\sqrt{N}} \sum_{k=1}^{N} |k\rangle|1\rangle \end{cases}$$

Partial Swap Test ($Q_{\text{1-query}}$)

- Want inner product: $\left(\sum_{j=1}^{N}\langle j|\langle E_{\theta j}|\right)\left(\sum_{k=1}^{N}|k\rangle|1\rangle\right) = \sum_{j=1}^{N} E_{\theta j}$

$$\text{Define } \begin{cases} |\phi_\theta\rangle \triangleq \frac{1}{\sqrt{N}}\sum_{j=1}^{N}|j\rangle|E_{\theta j}\rangle \\ |\psi\rangle \triangleq \frac{1}{\sqrt{N}}\sum_{k=1}^{N}|k\rangle|1\rangle \end{cases}$$

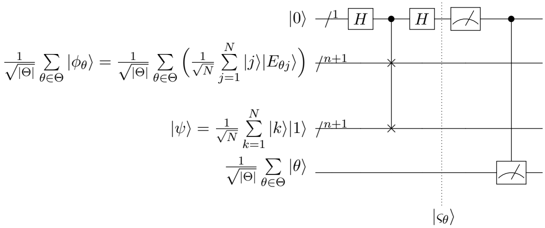


(1)

Partial Swap Test ($Q_{\text{1-query}}$)

- Want inner product: $\left(\sum_{j=1}^{N}\langle j|\langle E_{\theta j}|\right)\left(\sum_{k=1}^{N}|k\rangle|1\rangle\right) = \sum_{j=1}^{N} E_{\theta j}$

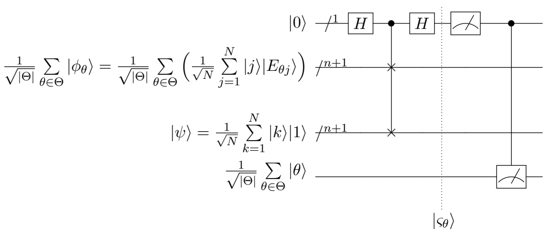
$$\text{Define} \begin{cases} |\phi_\theta\rangle \triangleq \frac{1}{\sqrt{N}}\sum_{j=1}^{N}|j\rangle|E_{\theta j}\rangle \\ |\psi\rangle \triangleq \frac{1}{\sqrt{N}}\sum_{k=1}^{N}|k\rangle|1\rangle \end{cases}$$



(1)

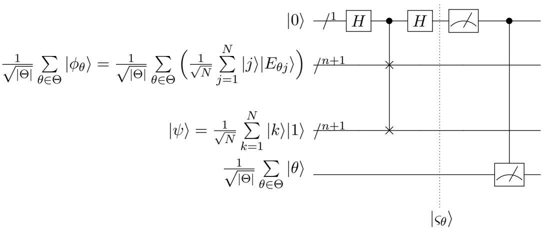
$$\boxed{\frac{1}{\sqrt{|\Theta|}}\sum_{\theta\in\Theta}|\varsigma_\theta\rangle|\theta\rangle = \frac{1}{\sqrt{|\Theta|}}\sum_{\theta\in\Theta}\frac{1}{2}|0\rangle\Big(|\phi_\theta\rangle|\psi\rangle + |\psi\rangle|\phi_\theta\rangle\Big)|\theta\rangle + \frac{1}{2}|1\rangle\Big(|\phi_\theta\rangle|\psi\rangle - |\psi\rangle|\phi_\theta\rangle\Big)|\theta\rangle}$$

Partial Swap Test ($Q_{\text{1-query}}$)



(2)

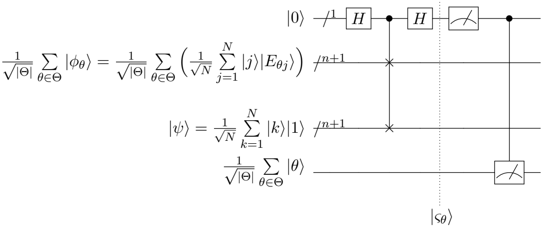
Partial Swap Test ($Q_{\text{1-query}}$)



(2)

$$\Longrightarrow \text{Lemma 2.2: } \mathbb{P}\Big[\text{1st qubit} = 0\Big] = \frac{1}{2} + \frac{1}{2|\Theta|} \sum_{\theta \in \Theta} |\langle \phi_\theta, \psi \rangle|^2$$

Partial Swap Test ($Q_{\text{1-query}}$)

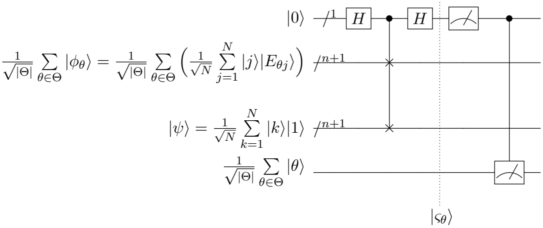


(2)

$$\Longrightarrow \text{Lemma 2.2: } \mathbb{P}\Big[\text{1st qubit} = 0\Big] = \frac{1}{2} + \frac{1}{2|\Theta|} \sum_{\theta \in \Theta} |\langle \phi_\theta, \psi \rangle|^2$$

$$\Longrightarrow \boxed{\text{Theorem 2.4: } \mathbb{P}\Big[\text{observe } \theta | \text{1st qubit} = 0\Big] \propto \frac{1}{2} + \frac{1}{2} \left| \frac{1}{N} \sum_{j=1}^{N} E_{\theta j} \right|^2} = \frac{1}{2} + \frac{1}{2} \left| \mathcal{L}_\theta^{\text{PSTC}} \right|^2$$

Partial Swap Test ($Q_{\text{1-query}}$)



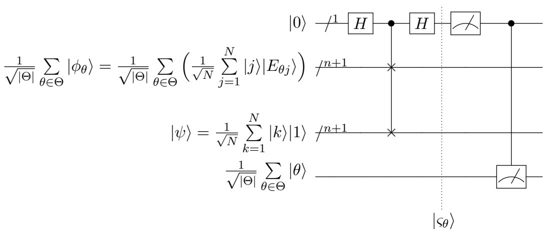
(2)

$$\Longrightarrow \text{Lemma 2.2: } \mathbb{P}\Big[\text{1st qubit} = 0\Big] = \frac{1}{2} + \frac{1}{2|\Theta|}\sum_{\theta\in\Theta}|\langle\phi_\theta,\psi\rangle|^2$$

$$\Longrightarrow \boxed{\text{Theorem 2.4: } \mathbb{P}\Big[\text{observe } \theta|\text{1st qubit} = 0\Big] \propto \frac{1}{2} + \frac{1}{2}\Big|\frac{1}{N}\sum_{j=1}^{N} E_{\theta j}\Big|^2} = \frac{1}{2} + \frac{1}{2}\Big|\mathcal{L}_\theta^{\text{PSTC}}\Big|^2$$

- Want larger $\mathcal{L}_\theta^{\text{PSTC}}$

Partial Swap Test ($Q_{\text{1-query}}$)



(2)

$$\implies \text{Lemma 2.2: } \mathbb{P}\Big[\text{1st qubit} = 0\Big] = \frac{1}{2} + \frac{1}{2|\Theta|}\sum_{\theta\in\Theta}|\langle\phi_\theta,\psi\rangle|^2$$

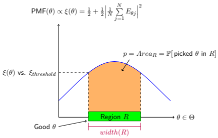
$$\implies \boxed{\text{Theorem 2.4: } \mathbb{P}\Big[\text{observe } \theta|\text{1st qubit} = 0\Big] \propto \frac{1}{2} + \frac{1}{2}\left|\frac{1}{N}\sum_{j=1}^{N}E_{\theta j}\right|^2} \quad = \frac{1}{2} + \frac{1}{2}\left|\mathcal{L}_\theta^{\text{PSTC}}\right|^2$$

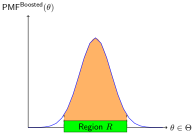
- Want larger $\mathcal{L}_\theta^{\text{PSTC}}$
- **The better θ we want, the higher chance we can observe θ!**

$\mathcal{A}_{\text{Boost}}$: Amplitude Amplification

$\mathcal{A}_{\text{Boost}}$: Amplitude Amplification

Table: Amplitude amplification for the PMF$(\theta) \propto \xi(\theta)$



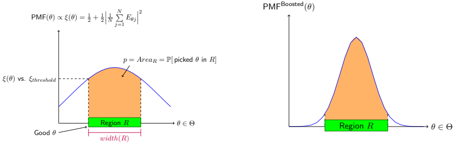


Before $\mathcal{A}_{\text{Boost}}$	After $\mathcal{A}_{\text{Boost}}$

$$p = Area_R = \mathbb{P}[\text{ picked } \theta \text{ in } R]$$

$\mathcal{A}_{\text{Boost}}$: Amplitude Amplification

Table: Amplitude amplification for the PMF$(\theta) \propto \xi(\theta)$



Before $\mathcal{A}_{\text{Boost}}$ After $\mathcal{A}_{\text{Boost}}$

$$p = Area_R = \mathbb{P}[\,\text{picked } \theta \text{ in } R]$$

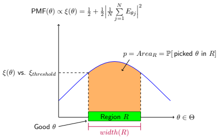
- In the manner $\sim$ Grover's Search [1]

$$\#(Q_{\text{1-query}}) = O(\tfrac{1}{p}) \quad \xmapsto{\;improved\;} \quad \#(Q_{\text{1-query}}) = O(\tfrac{1}{\sqrt{p}})$$

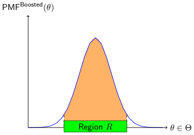
via classic algorithm via $\mathcal{A}_{\text{Boost}}$

$\mathcal{A}_{\text{Boost}}$: Amplitude Amplification

Table: Amplitude amplification for the $\text{PMF}(\theta) \propto \xi(\theta)$



Before $\mathcal{A}_{\text{Boost}}$



After $\mathcal{A}_{\text{Boost}}$

$$p = Area_R = \mathbb{P}[\,\text{picked } \theta \text{ in } R]$$

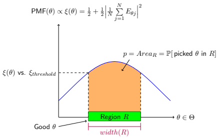
- In the manner $\sim$ Grover's Search [1]

$$\underset{\text{via classic algorithm}}{\#(Q_{\text{1-query}}) = O(\tfrac{1}{p})} \xmapsto{improved} \underset{\text{via } \mathcal{A}_{\text{Boost}}}{\#(Q_{\text{1-query}}) = O(\tfrac{1}{\sqrt{p}})}$$

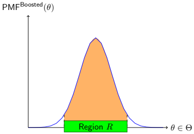
- Checking process: To determine whether $\theta \in R$

$\mathcal{A}_{\text{Boost}}$: Amplitude Amplification

Table: Amplitude amplification for the $\text{PMF}(\theta) \propto \xi(\theta)$



Before $\mathcal{A}_{\text{Boost}}$



After $\mathcal{A}_{\text{Boost}}$

$$p = Area_R = \mathbb{P}[\,\text{picked } \theta \text{ in } R]$$

- In the manner $\sim$ Grover's Search [1]

$$\#(Q_{\text{1-query}}) = O(\tfrac{1}{p}) \text{ via classic algorithm} \xmapsto{improved} \#(Q_{\text{1-query}}) = O(\tfrac{1}{\sqrt{p}}) \text{ via } \mathcal{A}_{\text{Boost}}$$

- Checking process: To determine whether $\theta \in R \leftarrow$ Cost $O(N)$

$\mathcal{A}_{\text{PSTM}}$

Lemma 2.5: $O(\frac{1}{\sqrt{p}}) \leq O(\sqrt{|\Theta|})$

$\mathcal{A}_{\text{PSTM}}$

Lemma 2.5: $O(\frac{1}{\sqrt{p}}) \leq O(\sqrt{|\Theta|})$

- Cost of $\mathcal{A}_{\text{Boost}}$: $O(\sqrt{|\Theta|}N)$

$\mathcal{A}_{\text{PSTM}}$

$$\text{Lemma 2.5: } O(\frac{1}{\sqrt{p}}) \leq O(\sqrt{|\Theta|})$$

- Cost of $\mathcal{A}_{\text{Boost}}$: $O(\sqrt{|\Theta|}N)$
- $\mathcal{A}_{\text{PSTC}}$: Run $\mathcal{A}_{\text{Boost}}$ for $O(\log|\Theta|)$ times to reduce $width(R)$

$\mathcal{A}_{\text{PSTM}}$

$$\text{Lemma 2.5: } O(\frac{1}{\sqrt{p}}) \leq O(\sqrt{|\Theta|})$$

- Cost of $\mathcal{A}_{\text{Boost}}$: $O(\sqrt{|\Theta|}N)$
- $\mathcal{A}_{\text{PSTC}}$: Run $\mathcal{A}_{\text{Boost}}$ for $O(\log|\Theta|)$ times to reduce $width(R)$
- Cost of $\mathcal{A}_{\text{PSTC}}$: Looks like $O(\sqrt{|\Theta|}N\log|\Theta|)$, but in fact $O(\sqrt{|\Theta|}N)$

$\mathcal{A}_{\text{PSTM}}$

$$\text{Lemma 2.5: } O(\frac{1}{\sqrt{p}}) \leq O(\sqrt{|\Theta|})$$

- Cost of $\mathcal{A}_{\text{Boost}}$: $O(\sqrt{|\Theta|}N)$
- $\mathcal{A}_{\text{PSTC}}$: Run $\mathcal{A}_{\text{Boost}}$ for $O(\log|\Theta|)$ times to reduce $width(R)$
- Cost of $\mathcal{A}_{\text{PSTC}}$: Looks like $O(\sqrt{|\Theta|}N\log|\Theta|)$, but in fact $O(\sqrt{|\Theta|}N)$
- $\because$ Region R in later iteration is narrower than the previous iterations

i.e. $Cost_{\text{Final iteration}}+...+Cost_{\text{1st iteration}} = O(\sqrt{|\Theta|}+\frac{\sqrt{|\Theta|}}{2}+\frac{\sqrt{|\Theta|}}{4}+...) = O(\sqrt{|\Theta|})$

Overview

Equivalency Discussion

- Are the 2 objective functions θ^*_{avg} and θ^*_{PSTC} equivalent?

Equivalency Discussion

- Are the 2 objective functions θ^*_{avg} and θ^*_{PSTC} equivalent?
- I.e., are there constants C, c such that on any sample set and any choice of ℓ,
$$c\theta^*_{PSTC} \leq \theta^*_{avg} \leq C\theta^*_{PSTC}.$$

Equivalency Discussion

- Suppose such c, C existed. Consider a sample set of size 3 as below: Let the parameters be $1, 2, ...C + 1$. Let $\ell_{threshold} = 2$ and suppose that for all $i \in 2, ...C$, we have the losses:

Equivalency Discussion

- Suppose such c, C existed. Consider a sample set of size 3 as below: Let the parameters be $1, 2, ...C+1$. Let $\ell_{threshold} = 2$ and suppose that for all $i \in 2, ...C$, we have the losses:
- $$\ell_{i,1} = 2.5, \ell_{i,2} = \ell_{i,3} = 2$$

  and suppose that,

  $$\ell_{1,1} = \ell_{1,2} = \ell_{1,3} = 1.9$$

  $$\ell_{C+1,1} = 2.1, \ell_{C+1,2} = 1, \ell_{C+1,3} = 0$$

Equivalency Discussion

- Suppose such c, C existed. Consider a sample set of size 3 as below: Let the parameters be $1, 2, ...C+1$. Let $\ell_{threshold} = 2$ and suppose that for all $i \in 2, ...C$, we have the losses:
-
$$\ell_{i,1} = 2.5, \ell_{i,2} = \ell_{i,3} = 2$$

  and suppose that,

$$\ell_{1,1} = \ell_{1,2} = \ell_{1,3} = 1.9$$

$$\ell_{C+1,1} = 2.1, \ell_{C+1,2} = 1, \ell_{C+1,3} = 0$$

- Here, $\theta^*_{avg} = \underset{\theta}{\operatorname{argmin}} \frac{1}{3} \sum_{j=1}^{3} \ell_{\theta j} = C + 1$, but

  $\theta^*_{PSTC} = \underset{\theta}{\operatorname{argmax}} \frac{1}{3} \sum_{j=1}^{3} \mathbb{1}[\,\ell_{\theta j} \leq \ell_{threshold}] = 1$.

Equivalency Discussion

- Suppose such c, C existed. Consider a sample set of size 3 as below: Let the parameters be $1, 2, ...C+1$. Let $\ell_{threshold} = 2$ and suppose that for all $i \in 2, ...C$, we have the losses:
- $$\ell_{i,1} = 2.5, \ell_{i,2} = \ell_{i,3} = 2$$

  and suppose that,

  $$\ell_{1,1} = \ell_{1,2} = \ell_{1,3} = 1.9$$

  $$\ell_{C+1,1} = 2.1, \ell_{C+1,2} = 1, \ell_{C+1,3} = 0$$

- Here, $\theta^*_{avg} = \underset{\theta}{\operatorname{argmin}} \frac{1}{3} \sum_{j=1}^{3} \ell_{\theta j} = C+1$, but

  $\theta^*_{PSTC} = \underset{\theta}{\operatorname{argmax}} \frac{1}{3} \sum_{j=1}^{3} \mathbb{1}[\ell_{\theta j} \leq \ell_{threshold}] = 1$.
- So, $C+1 \leq C$ gives a contradiction.

Overview

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\|\Theta\|}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\|\Theta\|}N)$

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\lvert\Theta\rvert}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\lvert\Theta\rvert}N)$

- The overall cost of average approach is same as PSTC's

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\lvert\Theta\rvert}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\lvert\Theta\rvert}N)$

- The overall cost of average approach is same as PSTC's
- The cost of PSTC at quantum parallelism is much cheaper

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\lvert\Theta\rvert}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\lvert\Theta\rvert}N)$

- The overall cost of average approach is same as PSTC's
- The cost of PSTC at quantum parallelism is much cheaper
- Checking process cost (i.e. check whether $\theta \in R$) of PSTC is still expensive

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\lvert\Theta\rvert}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\lvert\Theta\rvert}N)$

- The overall cost of average approach is same as PSTC's
- The cost of PSTC at quantum parallelism is much cheaper
- Checking process cost (i.e. check whether $\theta \in R$) of PSTC is still expensive
- The "fundamental cost" of PSTC should be $O(\sqrt{|\Theta|} \cdot sublinear\ N))$

Summary

Approach	Quantum Parallelism	Checking Process	Overall
Average	$O(N)$	N / A	$O(\sqrt{\|\Theta\|}N)$
Cut-off (PSTC)	$O(1)$	$O(N)$	$O(\sqrt{\|\Theta\|}N)$

- The overall cost of average approach is same as PSTC's
- The cost of PSTC at quantum parallelism is much cheaper
- Checking process cost (i.e. check whether $\theta \in R$) of PSTC is still expensive
- The "fundamental cost" of PSTC should be $O(\sqrt{|\Theta|} \cdot sublinear\ N))$
- Future work: to lower the cost of PSTC to $O(\sqrt{|\Theta|} \cdot sublinear\ N)$ by enhancing the checking process

Overview

Additional Restriction

Additional Restriction

- In optimization problem, we would normally search θ over Θ as below:

$$\underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

Additional Restriction

- In optimization problem, we would normally search θ over Θ as below:

$$\underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

- It is also common for us to have additional condition on θ (e.g. fairness, smoothness, or regularization):

$$\underset{\theta \in \Theta \cap A}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

Additional Restriction

- In optimization problem, we would normally search θ over Θ as below:

$$\underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

- It is also common for us to have additional condition on θ (e.g. fairness, smoothness, or regularization):

$$\underset{\theta \in \Theta \cap A}{\operatorname{argmin}} \frac{1}{N} \sum_{j=1}^{N} \ell(\theta, x_j)$$

- In view of this, we can generalize the cut-off indicator $E_{\theta j}$ to ${E^{A}}_{\theta j}$:

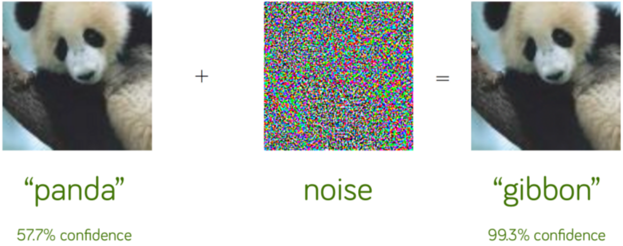
$$\mathcal{L}_{\theta}^{A,\mathsf{PSTC}} \triangleq \frac{1}{N} \sum_{j=1}^{N} {E^{A}}_{\theta j}, \text{ where } {E^{A}}_{\theta j} = \mathbb{1}[\,\ell(\theta, x_j) \leq \tilde{\ell} \text{ and } \theta \in A]$$

Adversarial Example

- The goal is to seek a perturbed input $x^{adv} = x + \delta^{adv}$
- The change δ^{adv} is not perceivable by human
- x^{adv} will be mis-classified by our target model h_θ

Adversarial Example

- The goal is to seek a perturbed input $x^{adv} = x + \delta^{adv}$
- The change δ^{adv} is not perceivable by human
- x^{adv} will be mis-classified by our target model h_θ



$$x + \delta^{adv} = x^{adv}$$

Adversarial Example

- One version of the adversarial problem can be formulated in this way:

$$x^{adv} = x + \delta^{adv} = x + \operatorname*{argmax}_{\|\delta\| \leq \epsilon} \ell(h_\theta(x + \delta), y)$$

Adversarial Example

- One version of the adversarial problem can be formulated in this way:

$$x^{adv} = x + \delta^{adv} = x + \underset{\|\delta\| \leq \epsilon}{\operatorname{argmax}} \, \ell(h_\theta(x + \delta), y)$$

- In this problem, the role of x and θ are switched: θ is now fixed, and x is the parameters to be optimized.

Adversarial Example

- One version of the adversarial problem can be formulated in this way:

$$x^{adv} = x + \delta^{adv} = x + \underset{\|\delta\| \le \epsilon}{\operatorname{argmax}} \, \ell(h_\theta(x+\delta), y)$$

- In this problem, the role of x and θ are switched: θ is now fixed, and x is the parameters to be optimized.
- Generalize our cut-off-qubit from $E_{\theta j}$ to $E_{\theta j}^{A} = E_{\theta j}^{\|\delta\| \le \epsilon}$, where $E_{\theta j}^{\|\delta\| \le \epsilon} = \mathbb{1}[\, \ell(\theta, x_j) \ge \tilde{\ell}$ and $\|\delta\| \le \epsilon]$

Adversarial Example

- One version of the adversarial problem can be formulated in this way:

$$x^{adv} = x + \delta^{adv} = x + \underset{\|\delta\| \leq \epsilon}{\operatorname{argmax}}\, \ell(h_\theta(x+\delta), y)$$

- In this problem, the role of x and θ are switched: θ is now fixed, and x is the parameters to be optimized.
- Generalize our cut-off-qubit from $E_{\theta j}$ to $E^A_{\theta j} = E^{\|\delta\| \leq \epsilon}_{\theta j}$, where $E^{\|\delta\| \leq \epsilon}_{\theta j} = \mathbb{1}[\,\ell(\theta, x_j) \geq \tilde{\ell}$ and $\|\delta\| \leq \epsilon]$
- Apply PSTC by searching for "good" $\delta \in X$ and we will be able to obtain an adversarial example in the cost of $O(\sqrt{|X|})$.

Thank you!

Overview

Notations Table: Basics

Notation	Space	Meaning
θ	Θ	Parameter of model $h_\theta(\cdot)$
$h_\theta(\cdot)$	$X \to Y$	Model with parameter θ
x	X	An input
y	Y	A label
S	$S \subseteq X$	$\{x_1, ..., x_N\}$
N	$\mathbb{Z}^+$	$\#$(samples in S)
n	$\mathbb{Z}^+$	$N = 2^n$
m	$\mathbb{Z}^+$	$\lvert\Theta\rvert = 2^m$
$\ell(\cdot,\cdot)$, or $loss(\cdot,\cdot)$	$\Theta \times X \to \mathbb{R}^+$	Loss function
	(or $Y \times Y \to \mathbb{R}^+$)	(depending on context)
ℓ_{ij}, or $\ell_{\theta_i j}$, or $\ell_{\theta j}$	$\mathbb{R}^+$	$\ell_{\theta_i j} = \ell(\theta_i, x_j) = \ell(h_{\theta_i}(x_j), y_j)$
$\mathcal{L}_\theta^{avg}$ or $\mathcal{L}^{avg}(\theta)$	$\mathbb{R}^+$	Average loss $\frac{1}{N}\sum_{j=1}^{N} \ell_{\theta j}$; The smaller the better
$\mathcal{L}_\theta^{\text{PSTC}}$ or $\mathcal{L}^{\text{PSTC}}(\theta)$	$\mathbb{R}^+$	Cut-off loss $\frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\ell_{\theta j} \le \tilde{\ell}]$; The larger the better
$\tilde{\ell}$, or $\ell_{threshold}$	$\mathbb{R}^+$	Threshold value for the cut-off approach
$E_{\theta j}$	$\{0,1\}$	Cut-off indicator: Value $\mathbb{1}[\ell_{\theta j} \le \tilde{\ell}]$ to be stored in 1 qubit
$\lvert\phi\rangle, \lvert\psi\rangle$	$\mathbb{C}^{n+1}$	Pure states used in PSTC

Notations Table: PSTC

$Q_{\text{1-query}}$	Q-Circuits	The Q-circuit for partial swap test
$\lvert\varsigma_\theta\rangle$	$\mathbb{C}^{2n+3}$	The pure state after the 2nd Hadamard gate in $Q_{\text{1-query}}$ given θ
$\lvert\varsigma_\theta^{(0)}\rangle, \lvert\varsigma_\theta^{(1)}\rangle$	$\mathbb{C}^{2n+2}$	The $\lvert 0\rangle$ and $\lvert 1\rangle$ parts of $\lvert\varsigma_\theta\rangle$
PMF$(\cdot)$	$\Theta \to [0,1]$	Probability mass function (PMF) of θ
$\mathcal{A}_{\text{1-query}}$, $\mathcal{A}_\xi$, $\mathcal{A}_{\text{Boost}}$, $\mathcal{A}_{\text{PSTC}}$	Algorithms	Names of some cut-off approach algorithms
$\theta_{\text{1-query}}$, θ_{Boost}, θ_{Best}	Θ	θ being used in the respective algorithms
$flag(\cdot)$	$\Theta \to \{0,1\}$	Flag to indicate whether θ is good or bad
O_{flag}	$\mathbb{C}^{m\times m}$	Gate of oracles on $flag(\theta)$
$\mathcal{R}$	$\mathbb{C}^{m\times m}$	Gate of "keeping zero, flipping else"
R	$R \subseteq \Theta$	Region R that "good" θs concentrate
$ref(\lvert B\rangle)$	$\mathbb{C}^{m\times m}$	Reflection on the "Bad" vector in $\mathcal{A}_{\text{Boost}}$
$ref(\lvert U\rangle)$	$\mathbb{C}^{m\times m}$	Reflection on the "Uniform" vector in $\mathcal{A}_{\text{Boost}}$
p	$[0,1] \in \mathbb{R}$	$\mathbb{P}[$picked θ in $R]$
$\xi(\cdot)$	$\Theta \to \mathbb{R}^+$	A function that proportional to pdf of θ : $P[$observe $\theta\rvert$1st qubit $=0]$
$\tilde{\xi}$, or $\xi_{threshold}$	$\mathbb{R}^+$	Threshold value of ξ for region R
$\lvert\mathbf{0}\rangle$	$\mathbb{C}^{\text{not care}}$	High dimensional $\lvert 0\rangle$ with dimension not mentioned
x^{adv}	X	Adversarial example
δ^{adv}	X	Perturbation
$\lvert\chi^{\text{Type I}}\rangle$, $\lvert\chi^{\text{Type II}}\rangle$	$\mathbb{C}^{\text{depends}}$	Pure states of x referring to Type I and Type II uniformity.

References

[1] R. de Wolf. Quantum computing: Lecture notes, 2019. URL https://arxiv.org/abs/1907.09415.

[2] C. Durr and P. Hoyer. A Quantum Algorithm for finding the Minimum. 1999. URL https://arxiv.org/abs/quant-ph/9607014.